格致方法
定量研究系列

· 革新研究理念
· 丰富研究工具
· 最权威、最前沿的定量研究方法指南

吴晓刚 / 主编
【美】威廉 · D. 贝里（William D. Berry） / 等著

# 线性回归分析基础



格致出版社 上海人民出版社

世纪出版

**格致方法**

定量研究系列

# 线性回归分析基础

本书由四种讨论定量方法的小册子组成，分别是《理解回归假设》、《回归诊断简介》、《虚拟变量回归》以及《多元回归中的交互作用》。本书的主要内容如书名所示，是介绍社会学研究分析方法之一，即线性回归。线性回归分析是社会科学研究中最常见的分析方法，本书通过介绍回归分析的假设，进而提出新的变量分析方法，对回归分析中的各变量及其相互关系进行阐述，为读者提供了一套完整的线性回归分析的方法。



上架建议：社科研究方法

ISBN 978-7-5432-1898-7

9 787543 218987 >

定价：58.00元

易文网：www.ewen.cc

格致网：www.hibooks.cn

格致方法
定量研究系列

吴晓刚 / 主编

【美】威廉·D. 贝里(William D. Berry) / 等著

# 线性回归分析基础

SAGE Publications ,Inc.

格致出版社 上海人民出版社

**图书在版编目(CIP)数据**

线性回归分析基础/(美)贝里(Berry, W. D.)等著;吴晓刚主编. —上海:格致出版社:上海人民出版社,2011

(格致方法·定量研究系列)

ISBN 978-7-5432-1898-7

Ⅰ. ①线… Ⅱ. ①贝… ②吴… Ⅲ. ①线性回归-回归分析 Ⅳ. ①O212.1

中国版本图书馆CIP数据核字(2010)第254750号

责任编辑 顾 悦

封面装帧 人马艺术工作室·储平

---

**线性回归分析基础**

[美]威廉·D. 贝里 等著

吴晓刚 主编

---

| | | |
|---|---|---|
| **出 版** | 世纪出版集团<br>www.ewen.cc | 格致出版社<br>www.hibooks.cn<br>上海人民出版社 |

(200001 上海福建中路193号24层)

编辑部热线 021-63914988

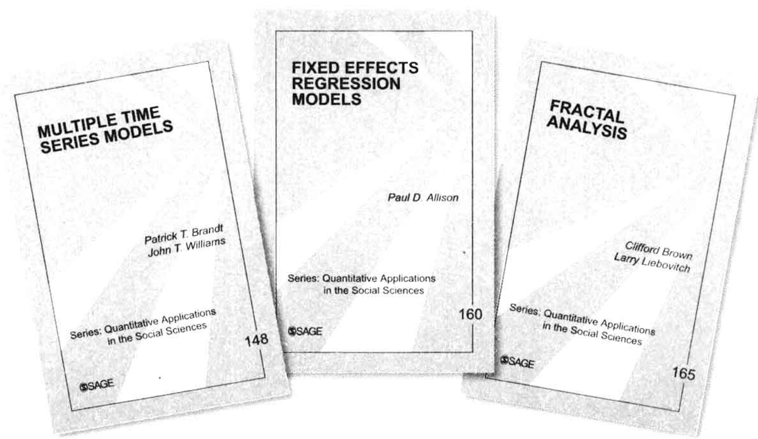
市场部热线 021-63914081

| | |
|---|---|
| **发 行** | 世纪出版集团发行中心 |
| **印 刷** | 上海江杨印刷厂 |
| **开 本** | 787×1092毫米 1/16 |
| **印 张** | 23 |
| **插 页** | 1 |
| **字 数** | 331,000 |
| **版 次** | 2011年7月第1版 |
| **印 次** | 2011年7月第1次印刷 |

ISBN 978-7-5432-1898-7/C·43

| | |
|---|---|
| **定 价** | 58.00元 |

# 绿皮书的传奇




1976 年，当政府管理与政治学教授 Eric Uslaner 向 SAGE 出版社的创始人 Sara 建议出版关于重要定量研究方法的简明小册子时，没有人预见到这套书会取得巨大成功。

那年夏天，“社会科学定量研究方法”丛书诞生。统一的、朴素的绿色封面，上面仅仅印着书名、作者名及本书所属的系列名。Iverson 和 Northporth 的《变量分析》(*Analysis of Variance*)、Nagel 的《运筹学》(*Operations Research*) 以及 Henkel 的《显著性检验》(*Tests of Significance*) 是最早出版的几本，每本售价 2.95 英镑，被形容为“像烤饼一样好卖”。

SAGE 选择了 20 种顶尖的定量研究工具，然后去寻找合适的作者，邀请他们围绕这些工具撰写 92 页的小书。这些薄薄的绿皮书在课堂上深受欢迎，在图书馆成套陈列。

至今，绿皮书系列中共有 160 种在初版或再版。书的主题反映了量化研究方法的发展：从基础统计知识、数据类型、测量到计算机的应用以及博弈论。这套书非常畅销，其中最畅销的一本，是 1980 年出版的 Michael Lewis-Beck 的《回归方法的应用》(*Applied Regression*)。

# 出版说明

本书由四种讨论定量方法的小册子组成，分别是《理解回归假设》、《回归诊断简介》、《虚拟变量回归》以及《多元回归中的交互作用》。本书的主要内容如书名所示，是介绍社会学研究分析方法之一，即线性回归。线性回归分析是社会科学研究中最常见的分析方法，该书通过介绍回归分析的假设，接着质疑假设，进而提出新的变量分析方法，最后对回归分析中的各变量及其相互关系进行阐述，为读者提供了一套完整的对线性回归分析的认识。因此，该书的问世能向社会科学研究者提供更深入的理论指导。

《理解回归假设》能使研究者深入了解多元回归分析的假设，并更熟练地驾驭回归分析，完成更有效的估计。《回归诊断简介》针对回归中经常出现的影响估计精度的因素，对研究者的假设提出质疑，运用“回归诊断”判断假设的合理性并处理回归分析中存在的问题。《虚拟变量回归》针对回归分析中，定序或名义变量无法有效反映因变量与自变量之间的实际关系，提出“虚拟变量”之概念，完善回归分析。最后，《多元回归中的交互作用》对最小二乘法中存在的交互作用项进行分析，厘清模型中各变量之间的关系和互相影响的情况，并提出了许多新的问题(如聚类数据的交互作用等)。

# 总序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫“调查数据分析”，第三门叫“研究设计”。除此以外，还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院（Institute for Social Research）定量社会研究方法项目的一些讨论会，受益良多。

2003年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课"社会科学里的统计学"(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课"社会科学中的定量分析"为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的"消费者"而非"生产者"。作为"消费者"，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。"社会科学里的统计学"(Statistics for Social Science)从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。"社会科学中的定量分析"则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的"定量社会科学研究"丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选

一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程，他们目前大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是：

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾“中央”研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景，书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异，校对也未免挂一漏万，术语的标准译法方面还有很大的改进空间，但所有的参与者都做了最大的努力，在繁忙的学习和研究之余，在不到一年的时间内，完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外，还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林，协助我通读了全稿，在此我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授，也参与了审校工作。

由于所选每本书都有一篇序言，对相关方法的背景和应用作了很好的介绍，我们均予以保留，内容在此不再赘述。为了方便起见，我们将内容相似的书目集册出版，每册三至五本不等，共八册，它们分别是：《线性回归分析基础》、

《高级回归分析》、《广义线性模型》、《列表数据分析》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》和《数据分析方法五种》。所冠书名未必能精准涵盖其中的内容，读者可自行参阅每本书的序言或目录。

我们希望本丛书的出版，能为推动国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

# 目 录

## 理解回归假设


序 3

第 1 章 简介 4

第 2 章 回归假设的正式描述 6

第 3 章 “体重”的案例 16

第 4 章 如何得到满意的回归假设结果 21

第 5 章 回归假设的实质意义 25

第 6 章 结论 79

注释 80

参考文献 85

译名对照表 87


## 回归诊断简介


序 91

第 1 章 概论 92

第 2 章 最小二乘回归 95

第 3 章 共线性 99

第 4 章 奇异值与强影响数据 109
第 5 章 非正态分布误差 126
第 6 章 不一致的误差方差 134
第 7 章 非线性 139
第 8 章 离散数据 145
第 9 章 最大似然法、计分检验和构造变量 150
第 10 章 建议 158
附录 163
参考文献 172
译名对照表 175

## 虚拟变量回归

序 179
第 1 章 简介 181
第 2 章 构建虚拟变量 187
第 3 章 虚拟变量回归 198

第 4 章　估计组影响差异 208
第 5 章　可替代虚拟变量编码方案 238
第 6 章　虚拟变量用法专题 249
第 7 章　结论 257
注释 258
参考文献 260
译名对照表 263

## 多元回归中的交互作用

序 267
第 1 章　导论 269
第 2 章　双向交互作用 283
第 3 章　三向交互作用 308
第 4 章　其他重要问题 323
注释 349
参考文献 350
译名对照表 353

# 理解回归假设

## 作者简介

**威廉·D. 贝里(William D. Berry)**

曾于美国佛罗里达州立大学和肯塔基大学讲授统计学和研究方法，现为佛罗里达州立大学政治科学系教授。其主要研究领域是公共政策和美国政策，在学术期刊上发表了大量论文，还参与撰写了《理解美国政府的成长：对战后时期的经验研究》(Praeger，1987)以及《实用多元回归》(Sage，1985)，同时他也是《非递归因果模型》(*Nonrecursive Causal Models*)(Sage，1984)一书的作者。

## 译者简介

**余珊珊**

2009年毕业于清华大学社会学系，目前是香港科技大学社会科学部硕士研究生。

# 序

回归分析是社会科学研究中最基本的工具,至少对于非经验主义者而言是这样。尽管它是一件最常用的工具,但它同样有可能是最容易被滥用的工具。每位一年级的研究生都会快速地学习构造最基本的多元回归模型,比如:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e$$

我们假设政治学家 Betty Brown 利用如下最小二乘估计模型(OLS)估计美国 50 个州的福利花费情况:

$$\hat{Y} = 543.66 + 87.10 X_1 + 450.39 X_2$$

其中 $\hat{Y}$ = 各州的福利花费(美元/人),$X_1$ = 民主党在国会里的议席(百分比),$X_2$ = 城市人口(百分比)。

Brown 教授可能会总结道,民主党的议席每增加 1%,福利花费的期望值就会增加87.1美元(当城市化水平保持恒定时)。那么这个对 $X_1$ 效果的估计到底有多好呢? 更确切地说,这是最好的线性无偏估计(BLUE)吗? 如果答案是肯定的,那么这一估计模型就能够与真实的世界联系起来。否则,这一估计模型只是那些流连在铅笔和草稿纸上的平面。

显然,我们应该去寻找能够达到最佳无偏估计标准的估计模型。这是我们学习回归假设的原因。Berry 教授非常严谨地定义了每一个假设,并且阐述了它们的实质意义。这种优美的文字描述搭配精选的图形和通俗易懂的证明,使得那些难懂的问题,比如测量、设定、多重共线性、异方差性以及自相关,都变得平易近人。而本书中的案例和数据也安排得很有条理,模型中的一个变量更能广泛地吸引人们的兴趣——体重。

理解回归假设可以让研究人员看到自己的弱点,同时也能够使他们更好地驾驭回归分析,以得到更有效的估计。当然,没有这种理解,就无法迈开通往构建模型的步伐。尽管目前已经有很多著作涉及回归分析这一话题,例如《应用回归》、《回归分析的解释和应用》、《实用多元回归》、《随机参数回归模型》、《理解回归分析》、《多元回归中的交互影响》、《回归诊断简介》,但是还没有人专门研究回归假设。Berry 教授的贡献恰好能填补这一空白。

迈克尔・S. 刘易斯-贝克(Michael S. Lewis-Beck)

# 第1章 | 简介

在任何回归分析被运用到社会科学研究中的时候,研究者总是会或明确或含蓄地提出无数的假设。[1]社会科学的定量研究已经非常流行,以至于几乎所有研究生二年级的学生都能够背诵一长串标准回归假设。然而尽管学生们经常死记硬背这些假设,却不能够理解其中"真正的含义"。多年来,我常常与研究生们针对他们的研究交换意见。而下文中所出现的屡见不鲜的场景正是让我决定撰写本书的原因:

教授:在你的模型中,你对异方差性这个概念还有问题吗(或者对任何其他的概念——设定残差、测量误差、自相关、非线性等等)?

学生:我不知道。

教授:那么,异方差性指的是什么?

学生(自信地):误差项的变化不是恒定的。

教授:好的。你的因变量是个人在慈善事业上的支出(或者任何其他变量)。你考虑了以下的自变量……在你的案例里面,如何解释误差项是异方差的?

学生(有点不自信了):对于不同的观测值,误差项的变异会有不同的取值。

教授:告诉我,这对于你的模型而言实质上意味着什么?你怎么解释慈善支出、你的模型中的自变量、其他影响慈善支出但没有包含在你的模型中的因素,以及所有这些变量是如何联系起来的?

学生(意识到自己知识上的一些漏洞被发现了):我真的不知道。

因此，尽管很多社会科学研究者能够自信地“不费吹灰之力地快速说出”一长串多元回归分析的假设（没有设定残差，没有测量误差，缺乏自相关等等），又或许他们能够说出这些回归假设的标准定义，但是常常缺乏对这些假设实质含义的深刻理解。如果我们对这些假设的理解仅仅局限于对定义的死记硬背，我们就无法把这些假设运用到对具体问题的分析中，这就相当于我们根本没有完全理解这些假设。

写作这本专题论著的目的是描述回归假设，并在某种程度上鼓励学生从死记硬背中解脱出来，转而去理解如何考察假设是否能够与一个具体的研究相适应。我们的讨论仅限于回归方法，因为回归在社会科学方法论中占据了主导地位，尽管也可以写出类似的关于其他的经验研究技术的著作。如果社会科学研究者能够仔细地考虑回归假设是否真正符合实际应用中的案例，而不是遇见什么问题都用回归方法来解决，那么当运用其他研究技术的时候，他们就能够更加自如地把握。

我以对标准多元回归假设的回顾作为开头，因为这些知识通常会出现在计量经济学或者回归分析的课本中。[2] 如果你不能理解这些假设的意义和重要性，不要担心。接下来，我会引入一个贯穿本书的具体案例，具体而言，这是一个关于体重的决定因素的模型。我选取这个案例是因为这里所涉及的人体的体重是与我们所有人都有关的话题——如果不考虑我们各自的兴趣——因此我们对此会有合理的直觉。最后，我回到回归假设，考察每一个假设的实际意义，并强调研究者如何评估每一个假设是否符合实际研究的需要。

# 第 2 章 | 回归假设的正式描述

## 回归分析概述

在标准多元回归模型中，假设因变量 $Y$ 是关于总体中 $k$ 个自变量（或者说是回归因子）$X_1$，$X_2$，$X_3$，…，$X_k$ 的方程。这个模型假设对于 $k$ 个自变量都有一系列取值——$X_{1j}$，$X_{2j}$，$X_{3j}$，…，$X_{kj}$——即关于 $Y$ 的条件概率分布，可以理解为（模型的）均值分布在"表面"上（有一个自变量的分布是一条直线；有两个自变量的分布是一个平面）。用以下的方程表示为：

$$E(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj}$$
$$= \alpha + \sum_{i=1}^{k} \beta_i X_{ij} \qquad [2.1]$$

在这一方程中，$Y_j$ 和 $X_{ij}$ 分别表示对第 $j$ 个观测值[3]，变量 $Y$ 和 $X_i$ 的取值。同时，符号"|"可以读成"给定"。所以 $E(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj})$ 表示在总体中当 $X_1 = X_{1j}$，$X_2 = X_{2j}$，…，$X_k = X_{kj}$ 时的均值或期望值。图 2.1 表示有两个自变量时的回归平面，

$$E(Y_j \mid X_{1j}, X_{2j}) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$$

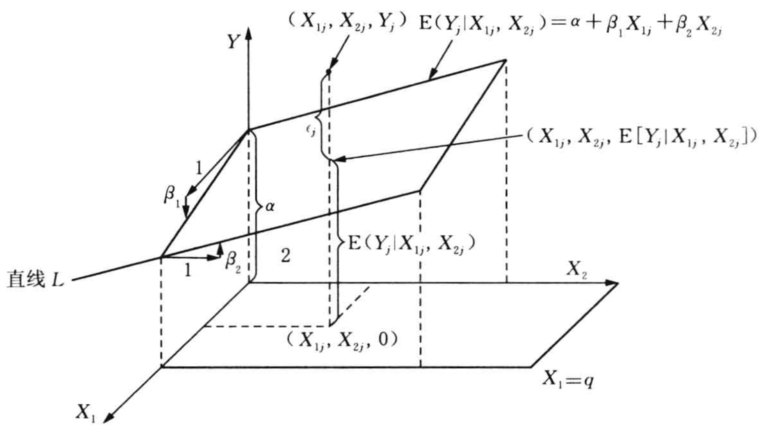
对只有一个自变量 $X$ 的双变量回归模型，我们可以对这一假设作出更详细的描述。图 2.2 表示对于 $X$ 的 3 个不同取值 $X_1$、$X_2$、$X_3$，$Y$ 的条件概率分布，其中纵轴表示概率。这样，分布的均值落在以下这条直线上：

$$E(Y_j \mid X_j) = \alpha + \beta X_j$$

这里所说的概率分布都被简化为正态分布并具有相同的方差。这一结论是从

回归假设 A8 和 A6 中得出来的，将在下文中进行分析。

回到方程 2.1 所表示的一般多元回归方程中，希腊字符系数（$\alpha$，$\beta_1$，$\beta_2$，…，$\beta_k$）表示总体系数或者参数。截距 $\alpha$ 表示当所有自变量全部等于 0 时，$Y$ 的期望值。在一般情况下，系数 $\beta$ 被称为“偏斜系数”。对于每一个 $i$，当所有其他自变量保持不变时，系数 $\beta_i$ 指的是当 $X_i$ 每变化一个单位的时候，$Y$ 的期望值随之变化的幅度。如果我们仅仅让回归统计量包含那些理论中提到的、对因变量有因果影响的自变量，$\beta_i$ 则可以被解释为 $X_i$ 对 $Y$ 的影响强度的测量值。在本书中，我将大量运用这种解释，即读者应该知道的是，在正式的多元回归假设中并不包括因果关系。因此，任何对于回归系数的“因果关系”的判断都必须基于回归分析以外的理论。




注：在这个三维空间中，每个点的位置是用 $X_1$、$X_2$ 以及 $Y$ 写在括号中，并且用逗号分开的值来表示的。举例而言，($a$, $b$, $c$)代表了 $X_1=a$, $X_2=b$, 以及 $Y=c$。

**图 2.1 有两个自变量的多元回归平面**

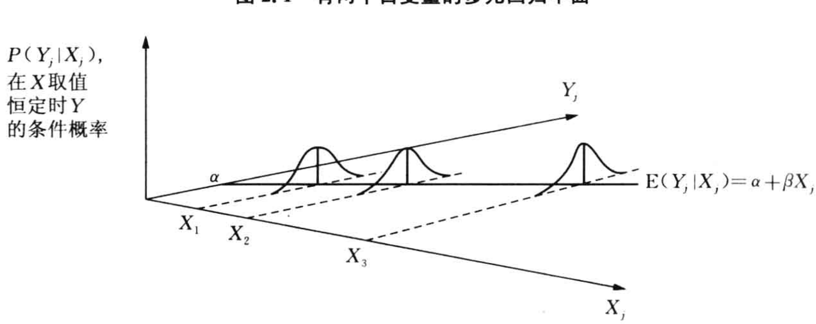



**图 2.2 二元模型的回归假设**

在解释 $\beta_i$ 的均值时，简单地说"当所有其他自变量都保持不变时"，而不明确指出具体哪些值保持不变，这样是说得通的。因为方程 2.1 的函数形式能说明这一点，所以自变量的作用是可叠加的。在一个可叠加的模型中，对于每一个自变量 $X_i$，当所有其他的自变量保持恒定时，$X_i$ 每增加一个单位，$Y$ 的期望值随之变化同样的幅度，而不论其他的 $X_i$ 的取值是否恒定。也就是说，每个自变量对因变量的作用不会根据其他自变量的取值而变化。在非叠加模型中，至少有两个变量交互作用并影响因变量。这就是说，如果一个自变量对 $Y$ 的影响随着其他自变量取值的不同而变化，那么就说明是两个变量交互作用来决定因变量 $Y$。

方程 2.1 的函数中也暗示了线性假设。当其他所有自变量保持恒定时，$X_i$ 被认为与 $Y$ 有线性关系。而 $Y$ 的期望值的变化与一小部分 $X_i$ 的增加幅度是一样的，不管 $X_i$ 具体取值是什么，$X_i$ 的斜率与 $Y$ 的期望值的关系是恒定的（如果 $X_i$ 与 $Y$ 的斜率随着 $X_i$ 取值不同而变化，$X_i$ 就被认为与 $Y$ 存在非线性的关系）。

对于回归系数的解释被反映在有两个自变量的图形中（图 2.1）。在图中，当 $X_1 = X_2 = 0$ 时，$\alpha$ 表示 $Y$ 的期望值，当 $X_2$ 保持不变时，$\beta_1$ 可以被看做由于 $X_1$ 增加了一个单位而导致 $Y$ 的期望值的变化。$\beta_2$ 则是当 $X_2$ 增加一个单位同时 $X_1$ 保持不变时，因变量期望值的变化。模型的线性特征反映在图形中，即回归表面是一个平滑的平面。由于回归表面是一个平面，如果我们让 $X_1$ 取 $q$ 值并保持不变，通过限制在垂直平面上 $X_1 = q$，并且让这个垂直平面与回归表面相交，那么得出的直线（或者称为线性的曲线）——用直线 $L$ 来表示——就有恒定的斜率 $\beta_2$。

## 误差项的作用

在回归模型中，尽管 $Y$ 的条件概率均值被假设为完全落在方程 2.1 的平面上，但 $Y$ 在每一个观测值 $j$ 上所取的真实值被认为是由自变量和误差项（或者说干扰项）$\varepsilon_j$ 共同决定的，如以下方程所示：

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + \varepsilon_j = \alpha + (\sum_{i=1}^{k} \beta_i X_{ij}) + \varepsilon_j \quad [2.2]$$

方程 2.2 还可以写成：

$$\varepsilon_j = Y_j - E(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj})$$

我们注意到，误差项 $\varepsilon_j$ 是 $Y$ 在观测值 $j$ 处所取的观测值 $Y_j$ 与 $Y$ 在总体模型中的预测值之差(即 $X_1 = X_{1j}$，$X_2 = X_{2j}$，$X_k = X_{kj}$)。图 2.1 中的图形显示，在有两个自变量的例子中，$\varepsilon_j$ 指的是 $Y$ 在 $j$ 点的取值与回归平面的垂直距离。想要理解回归假设的“实质意义”，理解误差项的含义是至关重要的。

为了理解误差项的含义，我们首先要考虑真实模型的含义。在研究回归分析的文献中，一个真实的模型通常被构想为一种可以解释总体中所有关于因变量的原因的模型。换言之，它可以完全解释因变量在总体中任何情况下取值的模型。因此这一方程反映了真实模型，包括所有可能影响因变量的变量，同时精确地反映了所有影响的性质。在具体的社会科学应用研究中，要想搞清楚真实模型是不可能的。但是，假设这样的模型存在或许是合理的，尽管对于研究者而言，这一模型实际上是不可知的。

当然，很多假设的基本前提是，因变量(在总体中)只能用一种方式解释，或者说只能用一个唯一的真实模型来解释。例如，Luskin 写道：

> 唯一的真实模型只存在于计量经济学的证明中。一个给定的因变量总是可以用很多等价且有效的方式进行解释——依靠大量的、理论得出的、异常细致的自变量或者基于少量的理论而略显粗糙的自变量，以及那些能够比较直接地反映其影响的变量，或者那些不直接地发挥其作用的变量。可能真的会有一个给定的独特的真实模型——在某种给定的概念集合上，或者给定的因果距离上。(Luskin, 1991:1038)

Luskin 对此作出了非常好的解释。例如，假设我们要解释在总体中，个人对一些具体事务的态度，我们可以构造方程来解释因果距离。在极端的近距离

下，我们能够通过其他更加普遍的态度来解释对于特定问题的态度。从一个更大的因果距离，我们可以发展出基于一个人的社会背景特征的模型；从远距离，我们可以构建一个依赖个人童年经验和社会化的模型。尽管有人可能争论说，一个真实模型需要在每个因果距离上合并变量，这种合并可能是不合适的。例如，如果有人可以通过一系列普遍性的问题来解释对一个具体问题的看法，同时结合社会背景特征来解释这些普遍的观点，那么这个用来解释人们对一个具体问题看法的模型，即用普遍性观点和社会背景特征这两种方法同时进行解释，是不合适的。取而代之的是，对于具体问题的态度，可能有两种不同的解释，而这两种不同的解释分别反映了不同的因果距离。

另一种解释是，即使在一个固定的因果距离上以及一个固定的概念集合体层次上，提出真实模型在社会科学研究的背景下也是没有用的。在这种观点下，不存在真实模型这种说法，而只存在理论。因此，试图评价一个回归模型是否符合一些"真实的"模型就是没有意义的。实际上，我们必须把自己限制在一种分析中，即对于那些影响到因变量的理论而言，一个回归模型是否精确地符合我们的理论。

我应该对"真实模型"做一些更加清晰的解释。事实上，对于能想到的每个社会科学因变量，我都怀疑是否存在一个真实的模型。即便真的存在这样的模型，我也怀疑我永远无法弄清楚这个模型。因此，在很大程度上，当研究开始进行时，思考这些模型是没有帮助的。另一方面，好的研究是由关键问题引导的，那是由理论和假设所激起的。与其担心回归模型是否符合某些假设的"真实"模型——那是我们永远无法知道的——我们更应该评判回归模型是否符合我们的理论，以及他们是否能够回答研究问题。不过，为了理解误差项的意义，假设存在一个未知的真实模型并把回归方程与假设模型进行比较，作为一种有启发意义的工具，还是很有用的。

所以，在这种情况下，假设在总体中有一种真实模型能够解释因变量 $Y$。毫无疑问，这将会是一个非常"长"的模型，因为这个模型会包括所有影响 $Y$ 的变量。假设其中一些变量对 $Y$ 有强烈的影响，但是也有其他的解释变量，它们对 $Y$ 的影响非常微弱。一种推理的路线是，真实模型会完全具有决定性的作用，这

也就是说，对于在总体内的任意情况，这会被算成对因变量的完美取值。[4] 在这种假设下，真实模型的形式是：

$$Y_j = \mu_0 + \mu_1 E_{1j} + \mu_2 E_{2j} + \cdots + \mu_p E_{pj} \qquad [2.3]$$

其中，$E_1$，$E_2$，…，$E_p$ 表示了一种有限——但是非常大——的解释变量的集合。[5]

但是绝大多数关于回归分析的教科书仍然采取这种观点，即使真实模型不是完全决定的，因为"有一些'内在的'变量会约束人的行为，以至于不能够被其他变量完全解释"(Gujarati, 1988:34；也可参见 Greene, 1990:144; Johnson, Johnson & Buse, 1987:43—44)。这些内在的随机性有时候是因为人类行为的"自由意愿"，或者完全是由"不可预知的事件"造成的(Kelejian & Oates, 1989:45)。一个悖论是，当所谓的人类行为的"内在的随机性"作为因变量的一部分时，我们会对它们进行更加准确的描述，但实际上并没有准备好如何去解释它们。从这种观点来看，把随机成分引入真实模型，意味着这个模型不再是"真实的"。在任何情况下，我们通过加入一个变量 $R$ 来对真实模型的最初方程进行修改。这一变量表示内在的随机成分。在 $Y$ 的行为中，得到：

$$Y_j = \mu_0 + \mu_1 E_{1j} + \mu_2 E_{2j} + \cdots + \mu_p E_{pj} + R_j \qquad \text{(真实模型)} \quad [2.4]$$

其中，$E$ 和 $R$ 一起能够完全说明 $Y$ 在总体中的差异(如果有人更倾向于前文中真实模型的确定版本，那么只能够假设 $R = 0$，因此把方程 2.4 转换为典型形式，即方程 2.3)。

在实际的研究中，我们永远不可能在经验分析中对真实模型进行研究。我们总是可以排除方程 2.4 中的一些(实际上是绝大部分)$E$，同时得到回归模型的假设，包括一个或多个自变量和一个误差项：

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + \varepsilon_j \qquad \text{(估计模型)} \quad [2.5]$$

为了说得更清楚，我们应该为这些来自真实模型的解释变量重新贴上标签。这些变量被从方程 2.5 的估计模型中剔除。新标签为 $Z$，因此能够清楚地区分被包括的变量($X$)和被剔除的变量($Z$)。于是，我们能够重新构造方程 2.4，把 $E$

分成$k$ 个$X$ 和$m$ 个$Z$(其中$k+m=p$,但是$k$ 比$m$ 小多了),同时对$Z$ 的偏斜系数重新命名:

$$Y_j = \mu_0 + (\beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj}) + (\delta_1 Z_{1j} + \delta_2 Z_{2j} + \cdots + \delta_m Z_{mj}) + R_j$$
$$= \mu_0 + (\sum_{i=1}^{k} \beta_i X_{ij}) + (\sum_{i=1}^{m} \delta_i Z_{ij}) + R_j$$

我们把$X$ 放到方程的一边,化简方程得到:

$$\sum_{i=1}^{k} \beta_i X_{ij} = (Y_j - \mu_0) - (\sum_{i=1}^{m} \delta_i Z_{ij}) - R_j \qquad [2.6]$$

接下来,我们重新构造假设模型 2.5,把误差项放在左边,得到:

$$\varepsilon_j = Y_j - \alpha - (\beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj}) = Y_j - \alpha - (\sum_{i=1}^{k} \beta_i X_{ij}) \qquad [2.7]$$

最后,从方程 2.6 中将表达式$\sum_{i=1}^{k} \beta_i X_{ij}$ 代入方程 2.7 中,可得到:

$$\varepsilon_i = Y_j - \alpha - Y_j + \mu_0 + (\sum_{i=1}^{m} \delta_i Z_{ij}) + R_j = (\mu_0 - \alpha) + (\sum_{i=1}^{m} \delta_i Z_{ij}) + R_j \qquad [2.8]$$

方程 2.8 表示,我们能够把一个回归模型中的误差项解释为,所有影响因变量的但却没有被包括在回归因子中的变量的联合作用,同时加上一些用任何在因变量中体现的内生的随机成分来表示的“随机的变量”。因此,假设我们永远无法把所有可能出现在真实模型中的变量包括进任何假设方程,所有的回归模型就必须包括一个误差项来解释这些被排除的变量的作用。

尽管很容易观察到,在任何具体的回归研究中,一个人总是会排除一些事实上会影响因变量的因素,但了解清楚排除这些因素的具体原因还是很有用的。首先,很多影响因变量的因素的实际作用是非常微弱的,因此,忽视这些变量则是明智的。另一方面,从表面来看,将所有对因变量有影响的变量都包含在模型中是一种很有道理的做法,即使这些因素对因变量的影响非常微弱。这是因为,如果这些变量被排除在外,某些被假设为对因变量仅有微弱影响,但实际上可能有更强影响的变量可能被忽略,且即使这个变量的作用非常微弱,把

这个变量包含在模型内，也需要用一个预先的假设来说明这个变量的作用是微弱的。但是，把一个对因变量仅有微弱作用的自变量包含在回归方程中是需要付出代价的。如果有微弱作用的变量与方程中其他有强烈影响的变量之间有高度的相关性，那么包括这些“微弱的”变量将会提高对这些作用“强烈的”变量的偏斜系数的估计效度。实际上，在很多情况下，如果所有影响因变量的因素都包括在回归方程里面，那么自变量的数量将会超过用于估计的样本的数量。我们将在后文中看到，这种情况将会导致完全多重共线性回归。这种情况违反了回归假设，同时使得从模型中无法得到有意义的偏斜系数的估计量。

第二，即使所有被认为对因变量有微弱影响的变量都被排除在回归模型之外，仍然会有足够的、对因变量有强烈影响的变量以及足够高的相关性在这些变量之间，来让估计量高度地不精确。如果是这样，在决定哪些对因变量有强烈影响的变量应该被包括进或排除出回归方程时，理论会变得很关键。假设不是所有对因变量有影响的解释变量都有相等的理论旨趣，例如，我们假定理论的目标是获得对 4 个变量 $X_1$、$X_2$、$X_3$ 和 $X_4$ 的影响的准确估计对于 $Y$ 的影响。如果是这样，最好排除那些与这 4 个变量仅有微弱联系的变量，而不是那些与这些 $X$ 有高度相关性的变量。

第三，将变量从一个回归中去掉的另一个原因是缺乏数据。很可能某些具体的变量在一些样本中本来就是无法被观察到的。同时，尽管我们并不情愿承认，但我们的研究选择有时候是由资源的可利用性决定的，比如，资金和时间的限制使得我们测量某些变量是很不实际的事情。同样的局限也有可能存在于数据收集阶段，即因为限制自变量的数量使得一些变量不得不被排除在外。

最后，还有被我们忽视的问题。即使一个变量对因变量有足够强大的影响，如果在理论推动下的回归模型没有指出这种解释变量的重要性，那么这一变量自然就会被淘汰。

在任何情况下，当考察关于回归假设的误差项是否符合实际应用的需要时，我们应该回到误差项的概念上，即所有影响因变量的，但是没有被包括在回归方程中的变量，加上任意的、内在的、影响因变量的随机因素。

## 其他回归假设

除了方程 2.2 中固有的假设形式，还有其他几种关于误差项的假设形式。因变量和自变量是标准回归模型中的形式：

A1. 所有的自变量 $X_1$, $X_2$, $X_3$, …, $X_k$ 是数量化的或者二分的，同时因变量 $Y$ 是数量化的、连续的以及无限的。[6] 而且，所有变量的测量都没有误差。

A2. 所有自变量都有非零方差（即每个自变量的取值都有一些差异）。

A3. 不存在完全多重共线性（即在两个或多个自变量之间没有完全的线性关系）。

A4. 在每一组 $k$ 个自变量中，$X_1$, $X_2$, $X_3$, …, $X_k$, $E(\varepsilon_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = 0$（即误差项的均值为 0）。

A5. 对任意一个 $X_i$，$COV(X_{ij}, \varepsilon_j) = 0$（即每个自变量与误差项都不相关）。[7]

A6. 在每一组 $k$ 个自变量中，$(X_1, X_2, X_3, \cdots, X_k)$，$VAR(\varepsilon_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = \sigma^2$，其中 $\sigma^2$ 是恒定的（即误差项的条件方差是恒定的）。这一条假设被称为“同方差性”。

A7. 对于任意两个观测值，$(X_{1j}, X_{2j}, X_{3j}, \cdots, X_{kj})$ 和 $(X_{1h}, X_{2h}, \cdots, X_{kh})$，$COV(\varepsilon_j, \varepsilon_h) = 0$（即不同观测值的误差项是不相关的）。这一条假设即缺乏自相关性。[8]

A8. 在每一组 $k$ 个自变量中，$\varepsilon_j$ 是正态分布的。

作为一组假设，A1 到 A7 被称为“高斯-马尔科夫假设”。

方程 2.2 是一个总体回归方程，其中的参数是未知的。然而，假设有来自于总体的样本数据，这些参数可以被估计。通常情况下，最小二乘估计法——或

者普通最小二乘法(OLS)估计量——是可以被确定的。[9]为了确保总体参数能够被清楚地与它们的估计量区分开,我将用 $a$ 表示 OLS 估计量的截距 $\alpha$,用 $b$ 来表示偏斜系数 $\beta_i$ 的估计量。

如果高斯-马尔科夫假设(A1 到 A7)都能被满足,那么最小二乘估计也有一些描述性特征(即无偏性和有效性),也可以被恰当地用于统计推论(比如,引入统计显著性,或者建立置信区间)。这些描述性特征将在下文中进行说明。

# 第 3 章 | “体重”的案例

为了说明回归假设的“实质含义”，在本书中，我将通过一个关于人体体重的例子来探索回归模型。这一案例包括 134 名年龄在 34 岁到 59 岁之间的女性。同时，我还假设我们已知真实模型的总体参数——因为这项研究不是在真实世界中展开的[10]：

$$\begin{aligned} \text{WEIGHT}_j = \alpha &+ \beta_C \text{CALORIES}_j + \beta_F \text{FAT}_j + \beta_E \text{EXERCISE}_j + \beta_H \text{HEIGHT}_j \\ &+ \beta_A \text{AGE}_j + \beta_S \text{SMOKER}_j + \beta_{FF} \text{FAT}_j^2 \\ &+ \beta_{SE}[(\text{SMOKER}_j)(\text{EXERCISE}_j)] \\ &+ \beta_M \text{METABOLISM}_j + \varepsilon_{Wj} \end{aligned} \quad [3.1]$$

本式满足所有的高斯-马尔科夫假设[11]，同时所有的变量有如下定义：WEIGHT(体重)指的是人体体重，单位为磅；CALORIES(卡路里)指的是前一年平均每天的食物摄入量，单位为卡路里；FAT(脂肪)指的是前一年平均每天饱和脂肪的摄入量，单位为克；EXERCISE(锻炼)指的是前一年平均每天通过体育活动消耗的能量，单位为卡路里；HEIGHT(身高)的单位为英尺，AGE(年龄)以周岁来衡量；SMOKER(吸烟)是一个二分变量，当样本是吸烟者时，此变量取 1，当样本不吸烟时，此变量取 0；METABOLISM(代谢率)指的是以“运气”[12]衡量的代谢率；$\varepsilon_W$表示误差项。总体参数如下所示：$\alpha$(截距) = 38.10；$\beta_C$(CALORIES) = 0.0291；$\beta_F$(FAT) =− 3.098；$\beta_E$(EXERCISE) = −0.1183；$\beta_H$(HEIGHT) = 1.346；$\beta_A$(AGE) =− 0.285；$\beta_A$(SMOKER) = 3.01；$\beta_{FF}$($\text{FAT}^2$) = 0.084；$\beta_{SE}$[(SMOKER)(EXERCISE)] = 0.1097；$\beta_M$(METABOLISM) =− 1.795。

在假设的总体中，有两个与体重有正向线性关系的决定性因素：身高和食物摄入量。身高的系数为 1.346，说明当所有其他变量保持恒定时，人的身高每

增加 1 英尺，预计体重相应地增加 1.35 磅。$\beta_C$的取值 0.0291 表示，当其他变量保持不变时，平均每天摄入的食物增加 100 卡路里，平均而言，体重会增加 2.91(＝100×0.0291)磅。我们注意到，当解释身高或食物摄入量的增加对体重的影响时，我们指的是横截面的“增加”，这也就是说，在总体中增加某个单位的某个变量的取值，而在本例中指的是样本中女性的某个变量取值增加。本例是把一位女性与另一位女性不同的身高或者食物摄入量相比较。在动态数据中，我们会更多地考虑“变化”，也就是说，我们会考察单个样本的某个变量取值的增减如何随时间的变化而发生改变。当引入时间序列回归时，即估计样本中包含对单个样本在不同时间点的观测值时，就能比较好地对偏斜系数进行动态解释。而利用横截面数据（在多个时间点上，能够反映某个特定时间点上样本之间关系的数据）时，也能够给出对偏斜率的动态解释。举例来说，有些人的兴趣是评估节食减肥的有效性——基于 $\beta_C$，这一系数是描述对女性而言，食物摄入量和体重的关系——也就是说，当其他变量保持恒定时，任何一个总体中的女性如果平均每天减少 100 卡路里食物消耗量，都预期能减重 2.91 磅。但是只有当第 2 章中的假设不再适用的时候，我们才会对回归模型进行调整，从现阶段的横截面回归转向跨时段回归分析。这些假设将在下文中进行讨论。

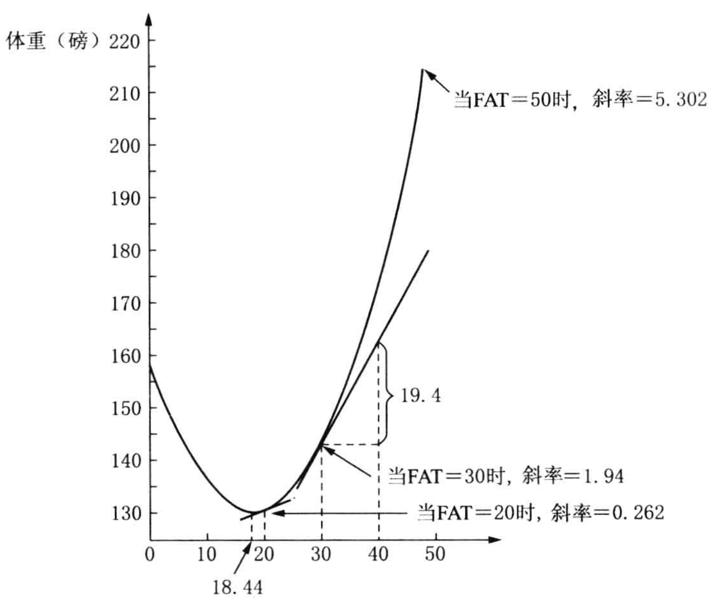
另外，在回归模型中，有两个决定因素与体重有负向的线性关系：AGE 和 METABOLISM。当其他变量保持不变时，年龄每增加 10 岁，预计体重相应减少 2.85(＝10×0.285)磅，而快速的代谢率会导致较轻的体重。具体而言，当其他变量保持恒定时，每增加一个单位“运气”的 METABOLISM，体重大概平均会减少 1.80 磅。

最后，作为自变量的脂肪摄入量与体重是非线性关系。其他两个变量——女性是否吸烟以及锻炼量——的交互作用会影响体重。这看上去可能有些奇怪，因为第 2 章介绍了线性和可叠加性是方程 2.2 固有的性质，但是一些非线性以及不可叠加性的模型，被认为实际上也是线性的和可叠加的。通过数学变换可以将这些模型转换为线性的和可叠加的，因此非线性以及不可叠加性都能符合标准 OLS 回归模型。我将在第 5 章中深入讨论非线性和不可叠加性。现在，我只是简单地解释方程 3.1 中的体重模型本质上是线性的和可叠加的。同时，我也会介绍对这

个模型中反映出来的线性和可叠加性的特殊性质进行解释。

正如全部的食物摄入量一样，饱和脂肪的消耗与体重之间也是正向的关系，但却并不是线性关系。正是 $FAT^2$（即变量 FAT 的平方）被包含到模型中才导致了非线性。图 3.1 展示了方程 3.1 中日均脂肪摄入量和预期体重的关系（此时方程中所有其他的变量都取总体均值）。在这个人为规定的“总体”中，饱和脂肪摄入量的取值范围大概从 20 克到 50 克。同时，这幅图也展现了，对于所有有意义的脂肪量的取值，饱和脂肪消耗量对体重的影响的强度随着消耗量的增加而递增。确实，从方程 3.1 中我们能得到，当所有其他自变量保持恒定时，在任意一个脂肪摄入量的固定取值上，即 $FAT^2$，表示脂肪和体重的期望值的关系的斜率为[13]：

$$\text{WEIGHT} = \beta_F + (2\beta_{FF} \cdot \text{FAT}^*) = -3.098 + [(2)(0.084)(\text{FAT}^*)]$$
$$= -3.098 + (0.168)(\text{FAT}^*) \qquad [3.2]$$




注：本图表示的是当总体中其他所有变量取它们的均值且保持不变时的情况（例如，CALORIES = 1.645，EXERCISE = 20.9，HEIGHT = 64.5，AGE = 46.8，SMOKER = 0.30，METABOLISM = 0 以及 SMOKER · EXERCISE = 5.3）。

**图 3.1　饱和脂肪摄入量与体重期望值之间的非线性关系（参见方程 3.1）**

所以,举例而言,当饱和脂肪摄入量为 20 克/天时,方程的斜率为 0.26[= 3.098 + (0.168)(20)],即当其他所有自变量都保持不变时,饱和脂肪摄入量每增加 1 克/天,会导致平均体重增加 0.26 磅。相较而言,当脂肪摄入量为 30 克/天时,同样的 1 克/天的增幅只会造成体重的期望值增加 1.94 磅。

方程 3.1 中的两个决定女性体重的且有交互作用的变量是:女性是否吸烟和体育运动量。回顾这些定义,当这两个变量交互作用并影响因变量 $Y$ 时,其中一个因变量对 $Y$ 的影响的强度依赖于另一个自变量的取值。[14] 方程 3.1 中的乘积项 SMOKER · EXERCISE 对这个模型产生交互作用。对于这个交互项,对于偏斜系数的典型解释是,当其他所有自变量的取值保持恒定时,自变量每增加一个单位,因变量随之变化的幅度。但这一解释并不能用到系数 $\beta_E$ 和 $\beta_S$ 上(它们分别为变量 EXERCISE 和 SMOKER 的系数)。为了对吸烟和锻炼的系数进行正确的解释,我们对方程 3.1 中的吸烟者和不吸烟者分别进行"评估"。对于不吸烟者,我们先"固定"吸烟为 0,即在方程中假设 $\text{SMOKER}_j = 0$。这意味着 SMOKER · EXERCISE 也为 0,那么所有这些回归量都被"剔除",方程可以被简化为:

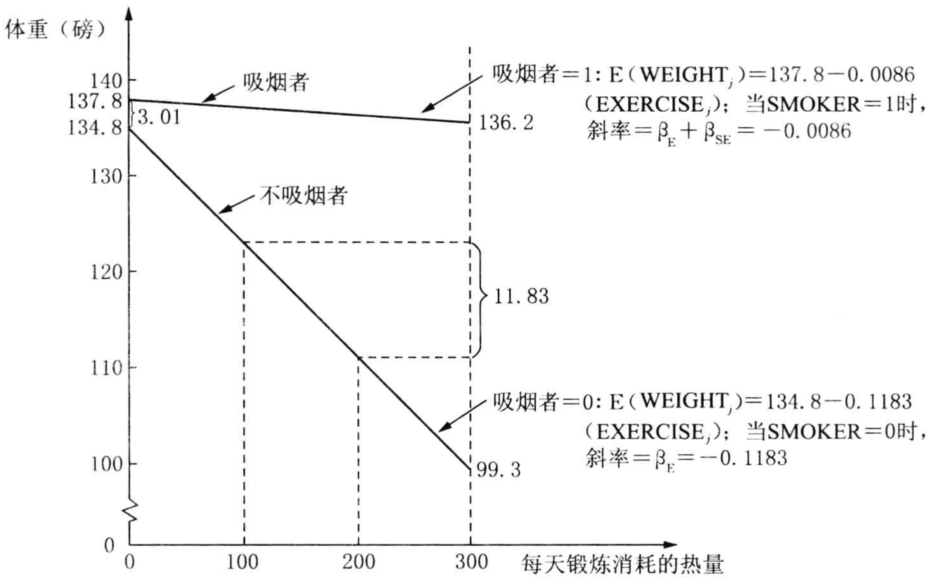
$$\text{WEIGHT}_j = \alpha + \beta_C\text{CALORIES}_j + \beta_F\text{FAT}_j + \beta_E\text{EXERCISE}_j + \beta_H\text{HEIGHT}_j + \beta_A\text{AGE}_j + \beta_{FF}\text{FAT}_j^2 + \beta_M\text{METABOLISM}_j + \varepsilon_{Wj} \quad [3.3]$$

接着对方程 3.1 中的吸烟者进行评估,设 SMOKER = 1,化简合并各项,可得:

$$\text{WEIGHT}_j = (\alpha + \beta_S) + \beta_C\text{CALORIES}_j + \beta_F\text{FAT}_j + (\beta_E + \beta_{SE})\text{EXERCISE}_j + \beta_H\text{HEIGHT}_j + \beta_A\text{AGE}_j + \beta_{FF}\text{FAT}_j^2 + \beta_M\text{METABOLISM}_j + \varepsilon_{Wj} \quad [3.4]$$

方程 3.3 和方程 3.4 表明,当其他自变量保持恒定时,方程 3.1 所反映的体重模型表示,运动量与体重的期望值在吸烟者和不吸烟者之间的关系是不一样的。这一点在图 3.2 中也反映出来了。对于不吸烟者,斜率为 $\beta_E$(−0.1183);而对于吸烟者,斜率则是 $\beta_E + \beta_{SE}$(−0.1183 + 0.1097 = −0.0086)。因此,对于不吸烟者,当其他自变量都保持不变时,通过剧烈运动消耗的能量每增加 100 卡路里/天,

将会导致 WEIGHT 期望值减少 11.83(= 100 · $\beta_E$) 磅。但对于吸烟者,同样的运动量只会导致 0.86(= 100 · [$\beta_E + \beta_{SE}$]) 磅体重的下降。注意,方程 3.3 和方程 3.4 的截距因为 $\beta_S$(3.01)而不同,这意味着当其他的自变量取任意固定值时,没有进行体育锻炼的吸烟者(即 EXERCISE = 0),其体重比不吸烟也没有进行剧烈运动的人要平均超出 3.01 磅。




注:图中所示的直线反映的是在总体中所有其他变量都取均值且保持固定的情况(例如,CALORIES = 1.645,HEIGHT = 64.5,AGE = 46.8,METABOLISM = 0 以及 FAT = 25.2)。然而,这两条直线的斜率并不随着其他取值已经固定的变量而变化,只有两者的交互项会随之变化。

**图 3.2 女性是否吸烟与锻炼量对体重的期望值的交互影响(参见方程 3.1)**

# 第 4 章 | 如何得到满意的回归假设结果

如果假设 A1 到 A7(即除误差项为正态分布之外的所有假设)都满足,高斯-马尔科夫理论保证了对 OLS 回归模型系数的估计量有两个理想的特征:无偏和有效(Berry & Feldman, 1985:15; Hanushek & Jackson, 1977:46—47; Johnson et al. , 1987:51; Wonnacott & Wonnacott, 1979:27)。[15]无偏性具有极其重要的意义,但是经常被错误地理解。对一个估计量 $\theta$(其总体参数为 $\theta$)和总体参数 $\theta$ 的估计量 $\hat{\theta}$,如果它的均值在重复随机取无限多的样本后等于被估计的参数,即 $E(\hat{\theta}) = \theta$,那么就称 $\hat{\theta}$ 为无偏的。此外,对于 $\theta$ 的无偏估计量 $\hat{\theta}$,如果在特定的无偏估计量中具有最小方差,则称之为"有效的"。所以,在图 4.1 中展现的 4 个概率分布(其估计量分别为 $\hat{\theta}_1$、$\hat{\theta}_2$、$\hat{\theta}_3$、$\hat{\theta}_4$),$\hat{\theta}_2$ 和 $\hat{\theta}_4$ 都位于总体参数 $\theta$ 的"中心",因此它们是无偏的。估计量 $\hat{\theta}_1$ 是负向有偏的,因为 $E(\hat{\theta}_1) - \theta < 0$,同时估计量 $\hat{\theta}_3$ 是正向有偏的。在这两个无偏估计量中,$\hat{\theta}_2$ 和 $\hat{\theta}_4$ 是有效的,因为它的方差最小,或者说,它是最精确的。[16]

高斯-马尔科夫假设(A1 到 A7)保证了回归系数估计量平方的最小值是无偏的。也就是说,OLS 估计量在样本取到无限大的时候要"命中靶心"。但是 OLS 估计量的这一特征无法保证每个基于单个样本回归总体的单个估计量能够与总体取值相等。相反,从总体中不断重复抽样能够对每个参数都产生一个概率分布估计——称为"抽样分布"——这一分布的均值即总体参数,但仍会包括比总体参数大或小的取值。[17]并且,在那些线性的、无偏的、有效的估计量中,回归参数的最小二乘估计量一定是有最小方差的抽样分布。用速记符号来表示,OLS 系数的估计量被描述为 BLUE——最优(表示有最小的抽样方差)、

线性且无偏的估计量。

误差项是正态分布的(A8)这一假设的重要性主要在于,它提供了关于回归系数估计量抽样分布"形状"的信息。除了假设 A1 到 A7 以外,当正态分布的误差项保持不变时,每个 OLS 回归系数的估计量的抽样分布也是正态分布的。因此,如果假设 A1 到 A8 都能满足,OLS 估计量 $b_i$(参数为 $\beta_i$)就能够精确地反映图 4.1 中 $\hat{\theta}_2$ 的正态概率分布。

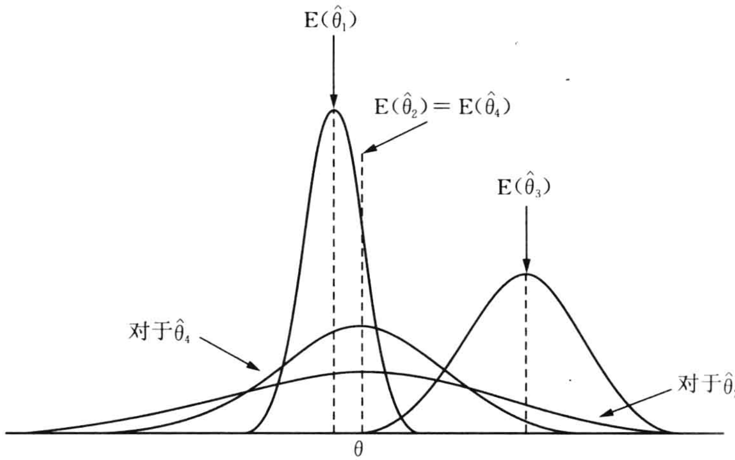



**图 4.1　4 种不同的对于参数 θ 估计量的概率分布**

对于数量较大的随机样本(确实是无限大的样本),其最小二乘估计量的期望值具有无偏性的特征,而对于单个的估计量则没有这一特性。为了说明这一点,我从 134 个虚构总体中的 50 个样本中随机抽取了 200 个身高样本,并利用 OLS 估计方程 3.1 中的参数。表 4.1 比较了 200 个样本的 OLS 每个平均回归系数的估计量(第二列)和真实总体的回归系数(第一列)。对每个参数,OLS 的平均估计量与总体值非常接近。确实,从第三列可以看出,平均估计量与参数值的比率一致地接近于 1.00,跨度从 0.945 到 1.112。但要注意,基于一个样本的估计量有巨大的潜在偏误。对于每个参数,从 200 个样本中得出的最小二乘估计量的最大值和最小值已经分别列在第四列和第五列。另外,图 4.2 展示了 200 个身高系数估计量 $\beta_H$ 的频率分布。

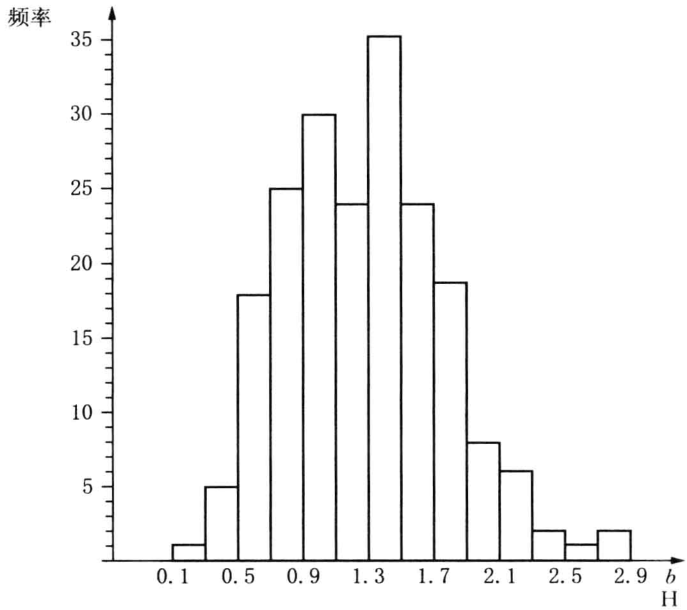



**图 4.2　体重模型中 200 个身高系数的估计量的分布**

**表 4.1　总体方程 3.1 的参数 OLS 估计的比较**

| 参数 | (1) 总体取值[a] | (2) 平均 OLS 估计量[b] | (3) 第二列/第一列 | (4) 最小 OLS 估计量[b] | (5) 最大 OLS 估计量[b] |
|---|---|---|---|---|---|
| $\alpha$ | 30.10 | 40.69 | 1.068 | −30.20 | 118.68 |
| $\beta_C$ | 0.0291 | 0.0295 | 1.013 | 0.0163 | 0.0449 |
| $\beta_F$ | −3.098 | −3.176 | 1.025 | −5.883 | −1.220 |
| $\beta_E$ | −0.1183 | −0.1315 | 1.112 | −0.2225 | −0.0208 |
| $\beta_H$ | 1.346 | 1.327 | 0.986 | 0.130 | 2.706 |
| $\beta_A$ | −0.285 | −0.300 | 1.053 | −0.575 | 0.092 |
| $\beta_S$ | 3.01 | 2.91 | 0.967 | −4.78 | 10.03 |
| $\beta_{FF}$ | 0.084 | 0.085 | 1.012 | 0.043 | 0.130 |
| $\beta_{SE}$ | 0.1097 | 0.1037 | 0.945 | −0.1298 | 0.3208 |
| $\beta_M$ | −1.795 | −1.814 | 1.011 | 2.456 | −1.244 |

注：a. 如正文所述。
b. 50 个样本中 200 次随机抽样。

这说明，基于单一的随机样本估计出来的 $\beta_H$ 不会非常精确（所得范围是小

于 0.130 或者大于 2.706)，但是却很有可能接近于分布的均值(1.327)。如果在上述频率分布中估计量的数量能够接近无限大，那么这个分布就会越来越接近正态分布。

记住，即使满足高斯-马尔科夫假设，对 OLS 回归模型的“最好的”线性无偏估计量而言，也不能保证它在所有的估计量中都是最好的——它只在线性无偏估计量中才是最好的。当然也不是说，所有无偏的估计量都比有偏的估计量要好。当然，在判断估计量的总体质量时，分析者需要同时考虑偏误和方差。举例而言，图 4.1 中的 $\hat{\theta}_2$ 是参数 $\theta$ 的一个无偏估计量，而 $\hat{\theta}_1$ 是有偏的，但 $\hat{\theta}_1$ 是更好的估计量。对于一个单独的样本，$\hat{\theta}_1$ 比 $\hat{\theta}_2$ 的值更有可能接近总体值 $\theta$。在接下来的章节中，我们将看到在 OLS 估计量中，几个由于违反了回归假设而导致偏误的情况。反之，有些回归量在违反回归假设时却仍然能够保持无偏，但对绝大部分无偏估计量而言，取得这样的精确性是要付出代价的。因此，当某个具体的回归假设被违背但 OLS 估计量仍然保持无偏时，我们并不能下结论说，违背回归假设是不重要的。从图 4.1 中的无偏估计量 $\hat{\theta}_1$ 的分布来看，当估计量有巨大差异而又无偏时，研究者不能心存侥幸。

# 第 5 章 | 回归假设的实质意义

## 从横截面回归中得出动态的解释

前文曾提到，回归模型可以是横截面的(指的是在同一时间点上，从各个不同角度观察样本)，或者是时间序列的(指的是在不同时间点上，只观察样本的一个方面)。[18]横截面回归模型的偏斜系数提供了很多信息，包括当自变量在一个单独的时间点上发生一定的变动(例如，出现差异)时，因变量的期望变化(或者差异)。相反，时间序列回归模型的系数告诉我们动态的或者跨时间段的变化，即它们会指出对于自变量从一个时间点到另一个时间点的变化，因变量的反应。由于时间序列回归所必需的数据通常无法得到，所以绝大部分社会科学回归分析都是横截面的。这些用来分析的数据通常包括个人、家庭、公司和其他组织、城市、国家或者民族。但是，在绝大多数情况下，能够动态解释的回归系数更加有趣。

举个例子，有一项研究旨在验证一个假设，即收入的增加会提高工作满意度。当我们说收入会影响工作满意度时，这通常指的是，如果我们选择一个人，并且调高他的收入，当保证其他变量恒定时，他(她)的工作满意度也会提高。一般来说，我们用横截面的方法来研究这一命题，因为收集一个人在不同时间段工作满意度的时间序列数据是非常困难的。然而，即使现实中的限制会让研究者转向利用横截面数据开展研究，我们仍然对个人收入和他(她)的工作满意程度的动态关系有着根本的兴趣。在公共政策分析中，人们主要的兴趣仍然在于利用动态数据进行解释的研究发现。例如，一项结论显示，A 政策相较 B 政策而言，大大提高了社区居民的生活质量。如果仅仅用横截面的数据来说明 A

政策比 B 政策对居民生活水平质量的提高的程度要高，那么这对政策制定者和学者而言，价值就非常小了。相反，如果辖区居民的生活水平因为政府由政策 A 转向政策 B 而大大提高，那么这个结论就有很大的理论和现实意义。

这并不是说横截面的关系对社会科学家而言是毫无意义的。如果真是这样，那些关于个人行为的理论永远不会将人类自身的特质——例如性别或者种族——作为自变量。因为这些变量的取值与其他变量不会有时间序列上的关系。确实，在回归模型中，如果对种族这一变量的偏斜系数进行动态解释，且因变量为政治党员身份，那么就会得出无意义的结论："当一个人的种族从白人变为非洲裔美国人时，他(她)的民主党党员身份的强度就会……"

不管怎样，在何种条件下对横截面回归模型的偏斜系数进行动态解释才是合适的呢？有两个假设必须满足。第一个是跨单位的不变性：决定一个样本的因变量跨时间段的取值范围的"过程"必须与上文中分析过的横截面中样本的取值过程是一致的。更正式地说，表达变量间横截面关系的回归方程必须精确地符合每个单位中因变量取值的过程。第二个假设是跨时间的不变性：任意一个单位的因变量取值变化的过程随着时间的推移，都必须保持稳定。每个自变量作用的级别(也就是每个偏斜率系数)必须保持一致。如果两个假设都得到满足，那么就假设我们有两个样本对于同一个自变量 $X_1$ 有不同的取值，$X_1'$ 和 $X_1''$，在同一个时间点给我们的信息必须与这一过程一致，即我们从一个样本在 $X_1$ 处的取值为 $X_1'$，过一段时间以后其取值变为 $X_1''$。因为跨单位和跨时间不变性的假设在任何实际应用中都不可能严格符合要求，所以实际的情况是这些假设都只能近似地满足。研究者应该抵制根据横截面回归系数进行动态解释，除非他们相信因变量的决定过程在时间和空间上都高度相似。

## 假设：缺乏完全多重共线性

在样本中，当观测值的自变量出现完全多重共线性时，会有无数回归平面(例如，在有两个自变量的例子中的回归平面)完全等价地"符合"因变量的观测值。因此，最小平方标准无法得出唯一的系数估计量。这种情况在研究设计中

很少出现，因为完全多重共线性要求在样本中，一个自变量（我们设其为 $X_i$）与其他自变量完全线性重合。这意味着 $X_1$ 可以表示为：

$$X_{ij} = c_0 + c_1X_{1j} + c_2X_{2j} + \cdots + c_{i-1}X_{i-1,\,j} + c_{i+1}X_{i+1,\,j} + \cdots + c_kX_{kj}$$

在这里，有部分但不是全部的常数 $c_0$，$c_1$，$c_2$，…，$c_{i-1}$，$c_{i+1}$，…，$c_k$ 有可能为 0。在这种情况下，如果用 $X_i$ 对剩下的自变量进行回归，那么 $R^2$ 将肯定为 1.00。

为了说明完全多重共线性的结果，我将验证一个只有两个自变量的模型：

$$Y_j = \alpha + \beta_1X_{1j} + \beta_2X_{2j} + \varepsilon_j$$

其中，$X_1$ 和 $X_2$ 有线性关系，如下面的方程所示：

$$X_{1j} = c + dX_{2j} \qquad [5.1]$$

其中，$\beta_2$ 可以被解释为，当 $X_1$ 保持不变时，$X_1$ 每增加一个单位，因变量 $Y$ 变化的期望值。但是如果 $X_1$ 和 $X_2$ 存在如方程 5.1 所示的联系，那么当 $X_1$ 保持不变时，要 $X_2$ 增加一个单位是不可能的。具体来说，$X_2$ 每增加 1 个单位，$X_1$ 就会增加 $d$ 个单位。基于这个原因，不可能在控制另一个自变量的情况下分离出一个自变量对因变量的影响。

实际上，完全多重共线性只会在 3 种情况下出现：第一，研究人员错误地将一系列"已经建立的"线性关系的因变量包括在内。例如，如果一位研究人员在方程 3.1 的自变量中加入出生年份（表示为 YEAR），那么就会在任何样本中出现完全多重共线性，对于所有个人，就有：

$$\text{年龄}_j = c - d\ \text{出生年份}_j$$

$c$ 表示现在的年份，同时 $d = 1$。一个类似的情况将会发生，如果一个分析人员在体重模型的自变量中加入"生活方式"指数，而且这个新变量是一个附加指数。这个附加指数是由卡路里消耗量（CALORIES）、剧烈运动消耗量（EXERCISE）以及他们是否吸烟（SMOKER）组成的。这一指数与模型中的其他 3 个自变量是完全的线性结合。

第二种可以导致完全多重共线性的情况是，当把离散的自变量纳入回归模型中时，处理虚拟变量可能出现的错误。当研究者用 $r$ 个虚拟变量来表示离散

变量的取值时，就会导致完全多重共线性。相反，实际上只有 $r-1$ 个虚拟变量能够包含在内。我们可以来看一个有太多虚拟变量的例子，这个例子包含了在回归方程中 3 种反映成人婚姻状况的分类：(1)现在已婚；(2)离婚或丧偶；(3)未婚。通过包含任意两个以下的二分变量，婚姻状态的影响可以在回归模型中很恰当地重组：

$$M_j = 1(\text{如果 } j \text{ 现在已婚，否则为 } 0)$$

$$D_j = 1(\text{如果 } j \text{ 离异或丧偶，否则为 } 0)$$

$$N_j = 1(\text{如果 } j \text{ 从未结过婚，否则为 } 0)$$

但是，如果分析人员把所有 3 个虚拟变量都包含在回归中，就会出现完全多重共线性。因为 3 个虚拟变量中的任意一个都与另外两个完全组合。

$$M_j = 1-(1)D_j-(1)N_j \qquad [5.2]$$

为了解释方程 5.2 的含义，我们首先认识到，婚姻状态变量可以区分 3 种类型的人，这些被分类的人以虚拟变量的取值形式出现在下表中：

| 类　型 | $M$ | $D$ | $N$ |
|---|---|---|---|
| 现在已婚 | 1 | 0 | 0 |
| 从未结婚 | 0 | 0 | 1 |
| 离婚或丧偶 | 0 | 1 | 0 |

那么，在方程 5.2 中，我们能够确认 $M$、$D$ 和 $N$ 之间的关系满足所有 3 类人。对已婚者来说，方程 5.2 可以得出 $1=1-0-0$；对那些从未结婚者，可得出 $0=1-0-1$；对离婚和丧偶者，我们可以得出 $0=1-1-0$。

在以上讨论会导致完全多重共线性的情况中，请注意这个问题是与模型特殊性有关的，而不与用来估计的数据的性质有关。确实，在这些情况中，没有任何一种数据集——不管有多大——能够让 OLS 回归运作出唯一的参数估计量。计量经济学家所指的那些凭借对自身描述的模型也不能得出唯一的系数估计量，因为它们并没有被清晰地辨识。

第三种导致完全多重共线性情况的发生，不是因为模型无法辨识，而是因为用来估计的样本量太小。具体而言，在方程中，只要观测值的数量小于变量

(自变量和因变量)的数量,就会出现完全多重共线性。例如,方程 3.1 中尝试估计的体重模型包括 10 个变量,如果数据不足 10 个观测值,将无法得出唯一的参数估计。

对于有两个自变量的回归模型,有一种几何学解释可以帮助我们弄清楚为什么样本量太少会导致多重共线性。假设一位研究者希望用身高和卡路里两个预测变量来估计一个改动后的体重模型。总体回归方程在这个案例中将是一个平面(即表面光滑的平面)。在三维空间中,用坐标轴分别表示卡路里、身高和体重的取值。如果研究者试图只用两个案例来估计这个方程,数据将会显示为空间中的两点。因此,估计方程的任务将会是在三维空间中寻找最适合这两点的平面。通过这两点的直线将非常适合这两点,所以任何包含这条直线的无数个平面都会非常符合这两点,因此无数组系数估计量将使得回归方程的 $R^2$ 为 1.00。

一种普遍的误解是,自变量之间任意"精确的"关系(线性的或非线性的)会导致完全多重共线性。事实并非如此。实际上,如果方程中包含两个变量,即使其中一个是完全单调的(即按照一定顺序排列的)且非线性的,另一个变量的转换也不会导致完全多重共线性。例如,方程 3.1 中同时包括 FAT 和 $FAT^2$,也没有导致完全多重共线性。

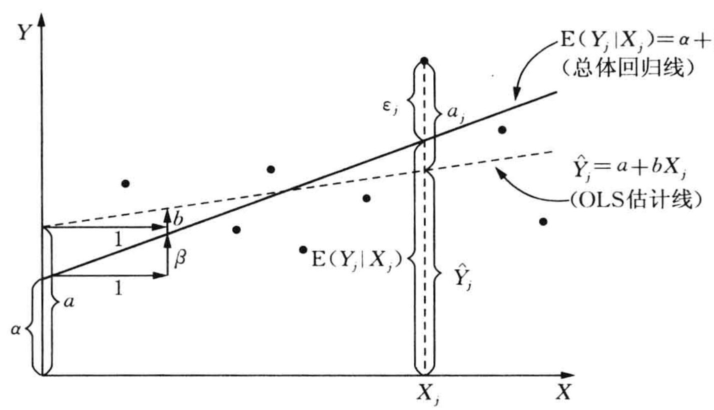
另一种流行的误解是,高度(但不是完全)多重共线性违反多元回归的假设。但是,回顾高斯-马尔科夫假设会发现,自变量之间近似的线性关系并不违背任何假设。因此,即使面对严重的多重共线性,OLS 参数假设仍然是最优无偏线性估计。这并不意味着分析人员没有为高度多重共线性付出代价。在这种条件下,对于那些具有共线性的变量而言,其偏斜系数估计量的标准误差将会非常高,因此,对于自变量的效果的估计将会在不同的样本之间有巨大的波动。[19]实际上,鉴于完全多重共线性通常被认为是鉴别的问题,强度较弱的多重共线性则被看做统计问题,但这一问题牵涉了估计样本中的众多自变量,而巨大的估计样本就很难对自变量的独特的效果作出精准的估计。

## 假设：误差项与每个自变量都没有相关关系

分析人员常常会把回归方程中误差项的含义和 OLS 回归残差弄混。残差被定义为，因变量的观测值与样本中基于最小二乘系数法估计得出的预测值的差异。更正式地说，如果方程 2.2 的系数由样本估计得出，并将这些估计量表示为 $a$, $b_1$, $b_2$, …, $b_k$，则任何回归残差 $j$——表示为 $e_j$——就被定义为：

$$e_j = Y_j - \hat{Y}_j = Y_j - (a + b_1 X_{1j} + b_2 X_{2j} + \cdots + b_k X_{kj})$$

其中，$\hat{Y}_j$ 表示基于样本系数估计量得出的 $Y$ 的预测值。在上图所示的估计样本回归表面中，$Y$ 表示纵轴，样本的残差 $e_j$ 表示估计回归表面和 $Y$ 在该点的观测值的纵轴间的距离。对于二变量模型，这一距离在图 5.1 中表现为 $X$ 在观测值处的取值 $X_j$。相反，误差项 $\varepsilon_j$ 反映了无法观测的总体回归表面和根据样本得到的观测值 $Y$ 的纵轴距离，这在图 5.1 的二变量模型中也能表现出来。必须认识到，对于自变量之间关系的经验研究与基于样本的 OLS 回归并不能得出自变量与误差项不相关的假设（即第 2 章中列出的 A5 假设）是否合理的结论。这是因为无论误差项 $\varepsilon$ 的分布究竟如何，最小二乘法标准本身就保证了回归残差项总是与所有自变量完全不相关。




注：$e_j$ 表示对观测值 $j$ 的回归残差，$\varepsilon_j$ 表示误差项。

**图 5.1 二元回归模型中的回归残差与误差项**

为了理解这一假设的本质含义,即误差项与每个自变量都不相关(A5),我们回到误差项的概念(可参见方程 2.8)即所有影响因变量但是被排除在回归方程之外的变量影响的集合,以及对因变量产生影响的任意随机因素。我们可以正式地写为:

$$\varepsilon_j = \delta_0 + \delta_1 Z_{1j} + \delta_2 Z_{2j} + \cdots + \delta_m Z_{mj} + R_j \qquad [5.3]$$

被排除的自变量标记为 $Z$, $R$ 表示随机成分,方程 2.8 中的截距被重新命名为 $\delta_0$。给定概念 $\varepsilon$,误差项与每个自变量都不相关的假设需要满足 $\sum_{i=1}^{m} \delta_i Z_{ij}$ 与每个包含在内的 $X$ 变量都不相关[20],尽管被排除的解释变量的线性组合(即 $\sum_{i=1}^{m} \delta_i Z_{ij}$)很有可能与每个自变量都不相关。如果自变量 $Z$ 不是那种基于理论的、在那些被排除在外的变量中与它的联系最紧密但是与其他被包含在内的变量联系最弱的变量,那么我们就有自信说假设 A5 是有道理的。

存在一种情况是,误差项与每个在回归方程中的自变量都不相关这种假设,当出现互为因果关系时,即当因变量影响一个或多个自变量时,肯定会被违反。例如,假设

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + \varepsilon_j \qquad [5.4]$$

所有高斯-马尔科夫假设都被满足,除了没有关于 $COV(X_{1j}, \varepsilon_j)$ 的假设。但是假定 $Y$ 是 $X_1$ 的一个原因。具体来说,假设

$$X_{1j} = \alpha^* + \beta^* \gamma_j + u_j \qquad [5.5]$$

其中,$\beta^* \neq 0$ 且 $u$ 为误差项。在这个例子里,$COV(X_{1j}, \varepsilon_j)$ 一定不为 0。本质上来说,因为 $\varepsilon$ 是一个影响 $Y$ 的误差项,所以当 $Y$ 转而影响 $X_1$ 时,$\varepsilon$ 是 $X_1$ 的一个“间接”原因。因此,$X_1$ 和 $\varepsilon$ 必然相关。[21]一种正式的证明是,当 $Y$ 是 $X_1$ 的一个原因时,要求证 $COV(X_{1j}, \varepsilon_j) \neq 0$ 需要更多的数学运算(参见 Gujarati, 1988:563—564,有与本例类似的模型的证明)。

## 设定误差:使用错误的自变量

在绝大多数情况下,当回归方程或者其中一个假设出现任何形式的错误

时，我们就认为出现了设定误差。但是绝大多数社会科学家都把这一术语用在更加狭义的范围里，即用它去描述变量被纳入回归模型时所出现的错误。在本章的下一个部分中考虑到了一种设定误差，即“正确的”变量也被包含在内，但是模型没有以实用的（或者数学的）形式准确地表达变量之间的关系。本部分的主题是，当回归方程被“错误的”自变量估计出来时出现的设定误差的情况，即相关的变量被排除在模型外，而不相关的变量却被包含在内，或者两种情况兼而有之。

概念“排除相关变量”和“包括不相关的变量”给出了一条用来判断变量与模型的相关性的标准。在文献中，关于特殊误差的讨论通常提出由两个参考框架来评估：(1)一个真实的模型(在第 2 章中曾讨论过)；(2)由理论驱动的回归。对于前者，研究者对一个回归模型的判断是基于这个模型在多大程度上符合一个能解释因变量的真实模型的。就后者而言，回归分析的使用者被假定为必须有一套理论，而设定的精确程度取决于回归模型能够在多大程度上反映理论。我的观点是，以理论作为参考框架来判断模型的特殊性，比用一个难以捉摸的“真实模型”来判断更有道理。

首先，正如我们已知的，对于这一概念提出的一些挑战是，是否存在一个唯一的真实模型去解释任何给定的因变量？其次，即使我们接受存在真实模型这一假设，模型中的所有变量仍然是不可知的。因此，实际上我们几乎无法相信一个真实模型能提供一个唯一明确的、能够与估计出来的回归估计模型进行比较的参考框架。

另一种替换——把理论当做参考框架——确实也有自身的风险。最主要的是，它会制造一种“倒转”的诱惑：一开始设计一个回归模型，且只包括那些能够在现有的样本中测量出来的变量，结束的时候再“塑造”一种精确地包括那些变量的理论。这种方法可以保证设定误差在名义上可以被去除，但是这种去除完全是表面上的，实际上它还是存在的。因此，无论何时我们建立一个作为分析回归模型恰当性的合适的参考框架，其实质是，我们必须非常严肃地反思那些没能把某些特定变量或者某些关系代入错误的方程形式中的理论。然而，考虑到一位研究者已经对理论的架构给予了足够细致的考虑，所得出的理论应该是合理的。同时，基于已有的数据问题能够阻止对理论的修改。理论在判断估

计模型时能显现出最为适合的参考框架。

所以，在本书中，我规定了设定误差的概念，从而排除那种有一个或多个解释变量包括在理论中但被排除在估计模型之外的情况。在分析这种误差的后果时，我们假设的策略是，有一个类似于回归方程的形式化了的理论。例如，

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_4 X_{4j} + \beta_5 X_{5j} + \beta_6 X_{6j} + \varepsilon_j \text{（理论模型）} \quad [5.6]$$

这一方程包含干扰项 ε，表示所有那些能够影响 $Y$，但在理论中无法完全被确定的因素，以及任何本质上会对 $Y$ 产生作用的随机要素。我们假设方程 5.6 满足高斯-马尔科夫假设。实质上，这一假设表明，这一理论精确地反映了 $Y$ 被确定的过程。如果数据能够支持直接估计方程（这样的话，估计模型将与理论完全相同），我们就能够保证 OLS 估计量是无偏的。但是，相反，我们将建立一个类似于参考框架的方程。假设估计模型排除了一个或多个 $X$，接着得出结果。例如，假设 $X_1$ 和 $X_2$ 被排除，得出：

$$Y_j = \alpha + \beta_3 X_{3j} + \beta_4 X_{4j} + \beta_5 X_{5j} + \beta_6 X_{6j} + u_j \quad \text{（估计模型）} \quad [5.7]$$

其中，估计方程的误差项用 $u$ 表示，以区别于理论方程中的 ε。

理解这种类型排除法的影响的关键是，认识到方程 5.7 并没有完全忽略 $X_1$ 和 $X_2$ 对 $Y$ 的影响，它只是降低了对方程误差项的影响。换句话说，$u$ 既是 ε 也是 $X_1$ 和 $X_2$ 的函数：

$$u_j = \varepsilon_j + \beta_1 X_{1j} + \beta_2 X_{2j}$$

而 ε 被假设为与每个 $X_3$、$X_4$、$X_5$ 和 $X_6$ 都不相关（假设 A5），除非 $X_1$ 和 $X_2$ 与其他自变量都没有关系，$u$ 被假设为与 $X_3$、$X_4$、$X_5$ 和 $X_6$ 都相关，反对假设 A5。除非 $X_1$ 与 $X_2$ 都与其他自变量不相关，鉴于违反了假设 A5，$u$ 必然会被假设与 $X_3$、$X_4$、$X_5$ 和 $X_6$ 相关。这样，我们可以对等地理解由于排除变量导致设定误差造成的后果以及违反误差项与每个自变量都不相关的假设。因此，由排除变量而造成的设定误差的结果与违反回归假设——即误差项与每一个自变量都不相关——的影响是相同的。然而，非常重要的一点是，我们必须认识到所有下列设定误差

的结果都满足一个前提,即参考框架(即理论)都满足高斯-马尔科夫假设。例如,我们应该看到,如果 $X_1$ 与 $X_2$ 与其他的 $X_3$、$X_4$、$X_5$ 和 $X_6$ 是不相关的,即使出现特殊误差,方程 5.7 中的偏斜系数的 OLS 估计量仍是无偏的。但是这种无偏性在“真实的”理论中不一定会发生(即方程 5.6 满足高斯-马尔科夫假设)。

基于这一点,我们可以更加明确地从最简单的案例开始,研究由排除变量导致的设定误差。包括两个自变量的参考模型:

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \varepsilon_j \qquad \text{(参考框架)}$$

以及排除其中一个自变量的参考模型:

$$Y_j = \alpha + \beta_1 X_{1j} + u_j \qquad \text{(估计模型)}$$

我们假定参考模型满足高斯-马尔科夫假设,所以 $\beta_1$ 和 $\beta_2$ 都大于 0,在假设样本中 $X_1$ 与 $X_2$ 是正向相关的。直觉告诉我们,因为 $X_2$ 被排除在估计回归之外,且 $X_1$ 与 $X_2$ 正向相关,所以一些 $X_2$ 对 $Y$ 的正向影响会同时对 $X_1$ 起作用,导致对 $X_1$ 对 $Y$ 的影响有可能被高估。这种直觉被证明是正确的。当 $X_2$ 被忽略时,$\beta_1$ 的估计量出现正向偏误,而整体中 $X_2$ 对 $Y$ 的影响强度以及样本中 $X_1$ 和 $X_2$ 的关系强度决定了这一偏误的级别。$E(b_1) = \beta_1 + b_{21}\beta_2$,其中 $b_{21}$ 是斜率系数,来自所谓的辅助回归,也就是说,样本中 $X_2$ 对 $X_1$ 的回归。如果 $X_1$ 和 $X_2$ 之间的相关性是正向的,那么 $b_{21}$ 也是正向的。

排除法的设定误差的含义也可以在普通情况下精确地表达,只要参照模型满足高斯-马尔科夫假设(Deegan, 1976; Maddala, 1992:162—163; Rao & Miller, 1971)。[22]假设参考框架模型包含 $r$ 个变量($X_1$, $X_2$, …, $X_r$),但是估计模型仅包含前 $g$ 个变量。对包含在内的变量 $X_i(1 \leqslant i \leqslant j)$,其 OLS 估计量的偏误由以下方程表示:

$$E(b_i) = \beta_i + \sum_{j=g+1}^{r} b_{ji}\beta_j \qquad [5.8]$$

其中,$b_{ji}$ 是 $X_i$ 的偏斜系数,来自包含所有 $g$ 个变量($X_1$, $X_2$, …, $X_g$)的辅助回归 $X_j$。因此,对那些排除在外的变量($\beta_{g+1}$, $\beta_{g+2}$, …, $\beta_{g+r}$)来说,$b_j$ 的偏误是偏斜系数的加权总和。其中,每个 $\beta_j$ 的权重(即 $b_{ji}$)是一个来源于辅助回归的测

量值，也即当所有其他包括在内的变量保持不变时，$X_j$ 和 $X_i$ 之间关系强度的测量值。

我将用"体重"的例子来描述关于排除法的设定误差的意义。假设在方程 3.1 中排除 METABOLISM(在样本中无法测量个体的 METABOLISM)，那么，对误差的具体回归是 $u_j = \varepsilon_j - 1.795(\text{METABOLISM}_j)$。误差项中的 ε 与模型中每个自变量都不相关(假设 A5)，但是它与 METABOLISM 的关系如何解释呢？在实际的研究中，我们可以进行估计和猜测，但是并不能得到肯定的答案。然而，在我们设计的案例中，所有变量之间的相互关系都被呈现在表 5.1 中。请注意，METABOLISM 与剩下的自变量不是高度相关的，它与其他变量的二分

**表 5.1　134 个女性样本的方程 3.1 中各个独立随机变量的关系**

| | (1) CALORIES | (2) FAT | (3) EXERCISE | (4) HEIGHT | (5) AGE |
|---|---|---|---|---|---|
| CALORIES | 1.00 | | | | |
| FAT | 0.78 | 1.00 | | | |
| EXERCISE | −0.35 | −0.25 | 1.00 | | |
| HEIGHT | 0.25 | 0.24 | −0.31 | 1.00 | |
| AGE | 0.27 | 0.30 | −0.25 | 0.09 | 1.00 |
| SMOKER | 0.29 | 0.35 | −0.06 | 0.06 | 0.27 |
| $FAT^2$ | 0.77 | 0.98 | −0.22 | 0.23 | 0.33 |
| SMOKER · EXERCISE | 0.10 | 0.17 | 0.34 | 0.03 | 0.12 |
| METABOLISM | 0.32 | 0.25 | 0.00 | 0.18 | −0.32 |

| | (6) SMOKER | (7) $FAT^2$ | (8) SMOKER · EXERCISE | (9) METABOLISM | (10) $R^2$ |
|---|---|---|---|---|---|
| CALORIES | | | | | 0.66 |
| FAT | | | | | 0.97 |
| EXERCISE | | | | | 0.41 |
| HEIGHT | | | | | 0.16 |
| AGE | | | | | 0.35 |
| SMOKER | 1.00 | | | | 0.37 |
| $FAT^2$ | 0.18 | 1.00 | | | 0.97 |
| SMOKER · EXERCISE | 0.24 | 0.18 | 1.00 | | 0.41 |
| METABOLISM | 1.00 | 0.24 | 0.03 | 1.00 | 0.31 |

注：列(1)到列(9)的值是二元相关的；列(10)的值是在方程 3.1 中，表左侧的项相对于其余的项的 $R^2$ 值。

关系取值的跨度从 0.32(CALORIES 和 AGE)直到几近为 0。另外，当 METABOLISM 对在方程 3.1 中的所有其他自变量进行回归时，得出的 $R^2$ 只有 0.31。绝大多数分析人员大概会总结道，METABOLISM 与剩下的自变量不存在高度的共线性。

尽管如此，当回归因子中除去 METABOLISM 时，却导致某些系数估计值出现了实质性的偏误。[23] 表 5.2 的列(2)表明，当代谢被排除时，方程 3.1

**表 5.2 设定误差示例：方程 3.1 中被排除的变量**

| 变 量 | 参 数 | (1) 总体取值(无设定误差)[a] | (2) 排除 METABOLISM 无设定误差下参数的期望估计量 |
|---|---|---|---|
| 截距 | $\alpha$ | 38.10 | 65.83 |
| CALORIES | $\beta_C$ | 0.0291 | 0.0199 |
| FAT | $\beta_F$ | −3.098 | −3.001 |
| EXERCISE | $\beta_E$ | −0.1183 | −0.1384 |
| HEIGHT | $\beta_H$ | 1.346 | 0.817 |
| AGE | $\beta_A$ | −0.285 | 0.190 |
| SMOKER | $\beta_S$ | 3.01 | 1.46 |
| FAT² | $\beta_{FF}$ | 0.084 | 0.080 |
| SMOKER · EXERCISE | $\beta_{SE}$ | 0.1097 | 0.1299 |
| METABOLISM | $\beta_M$ | −1.795 | — |

| | (3) 设定误差导致的百分比偏差[b] | (4) 排除变量 METABOLISM 和 FAT 无设定误差下参数的期望估计量 | (5) 设定误差导致的百分比偏差[c] |
|---|---|---|---|
| 截距 | 0.73 | −26.84 | −1.70 |
| CALORIES | −0.32 | 0.0424 | 0.46 |
| FAT | −0.03 | — | — |
| EXERCISE | 0.17 | −0.0854 | −0.28 |
| HEIGHT | −0.39 | 1.154 | −0.14 |
| AGE | −1.66 | 0.418 | −2.47 |
| SMOKER | −0.51 | 5.43 | 0.78 |
| FAT² | −0.05 | — | — |
| SMOKER · EXERCISE | 0.18 | 0.1033 | −0.06 |
| METABOLISM | — | — | — |

注：a. 如正文所述。

b. (列(2)变量值−列(1)变量值)/列(1)变量值。

c. (列(4)变量值−列(1)变量值)/列(1)变量值。

中的参数估计量的期望值。列(3)给出了一个"相对的"估计量偏误测量值。它表明了当变量 METABOLISM 被排除在回归模型以外时,得到的估计量的期望值与没有设定误差且没有误差的期望值之间的差异的比例。对于年龄这一变量,设定误差问题是最严重的,因为其期望值 $b_A$ 的符号都被逆转了。实际的情况是,人的年龄每增加 10 岁,体重就会减少 2.85 磅。但是错误的模型却指出,年龄每增长 10 岁,体重的期望值反而会增加 1.9 磅。另外,还有 3 个偏斜系数估计量的偏误至少被低估了 30%,即 CALORIES、HEIGHT 和SMOKER。

方程 5.8 告诉我们,如果要从关于排除法的设定误差中了解偏误的具体程度,我们需要知道:(1)(在参考框架方程中)每个被排除的变量的偏斜系数;(2)每个被排除的变量在辅助回归中对于所有被包含的变量所得出的偏斜系数。实际上,前一种系数——作为总体取值——在本质上是不可知的。对于后一种系数,没有被排除的变量的数据也是无法计算的。然而,当仅有单个的自变量被排除时,偏误量(方程 5.8 中的 $\sum b_{ji}\beta_j$)就简化为一个乘积。同时,理论驱动的回归分析有时候会允许合理的、关于这个乘积符号的推论:正的或者负的。例如,假设一位分析人员关心的是,在某些模型中,排除 $X_2$ 对偏斜系数估计量 $b_1$ 的影响。这个理论所隐含的逻辑会产生一种关于 $\beta_2$ 正负号的预测。在很多情况下,研究者也会掌握一些信息,或者至少是辅助回归中的 $X_1$ 对于所有包含在内的变量 $X_2$ 的斜率系数 $b_{21}$ 的符号的直觉判断。也许 $X_1$ 和 $X_2$ 之间的关系在其他研究中已经被解释得很清楚了,因此,关于 $b_{21}$ 的符号问题的经验判断是可知的。在另一些情况下,研究者不得不依赖"有根据的推断"来判断 $X_1$ 和 $X_2$ 之间的二分关系的符号。相应的,关于 $\beta_2$ 和 $b_{21}$ 的符号的预测导致了对偏误方向的清晰的预测,因为偏误是这两个值的乘积。

假设我们不知道方程 3.1 中的总体参数。尽管理论上我们相信 METABOLISM 会影响个人的体重,但数据的缺乏迫使我们把 METABOLISM 从估计方程中排除出去。那么我们能够推导出哪些合情合理的、关于受到设定误差影响的参数估计量的偏误的论断?[24] 如果我们没有先验的理论(或者甚至是直觉),即当控制剩下的变量时,关于代谢率和某个特定的自变量之间的符号关系的理

论，那么自变量的系数估计量中所存在的偏误方向是不能被预测的。

但是，在某些样本中，我们可以对可能的偏误方向作出一个合理的预测。例如，请考虑食物摄取量的参数(CALORIES) $\beta_C$。利用方程 5.8，把 METABOLISM 当做单一的被排除的变量，我们可以得到：

$$E(b_C) = \beta_C + b_{MC}\beta_M \qquad [5.9]$$

其中，$b_{MC}$是辅助回归中 CALORIES 的偏斜系数。这一辅助回归方程描述了 METABOLISM 在方程 3.1 中对其他所有自变量的影响(除了变量 METABOLISM)。假设我们的理论提出 $\beta_M$是负的，基于代谢越快的人倾向于吃得越多的印象，那么我们可以总结出 $b_{MC}\beta_M$是负的。方程 5.9 又暗示，基于一个排除了变量 METABOLISM 的方程得出的估计量 $\beta_C$很可能会“太过负面”，也就是说，为了与真实的参数等价，$\beta_C$要加上一个负数。这类信息到底有多大的作用，取决于 $\beta_C$在特定的样本中的估计值的符号。如果 $b_C$是负的，那么我们通过估计得到的信息很有可能“太负向了”，以至于我们无法确定有可能出现的真实参数的符号到底是绝对值较小的负数还是正数。更加可能的一种情况是，如果 $b_C$在某些样本中是正的，这个估计量可能是“太过负向的”信息将会让我们得出这样的结论，即我们的假设很有可能低估了食物摄取量和体重之间的正向关系的强度。

当两个或者更多的自变量在参考框架模型中被排除在估计回归之外时，偏斜系数估计量中的偏误方向就更加难以预测了。请注意，一个样本在一个估计量中的偏误是对被排除的变量的偏斜系数的加权总和。其中，加权的正负依赖于样本中被排除的和包含在内的变量的关系的性质。因此，预测偏误方向的第一步就是预测方程 5.8 中 $r-g$ 个斜率系数 $\beta_j$ 和加权值 $b_{ji}$ 乘积的总和。只有在非常特殊的情况下，所有的 $r-g$ 个乘积的总和才可能被预测为正的，或者所有的都被预测为负的，才能使研究人员对偏误的方向作出明确的预测。相反，如果其中一些结果被认为是正的，而另一些被预测为负的，除非有人能够超越对结果的符号的预测而直接预测精确的结果，否则在总和上符号为正的结果是否会“抵消”符号为负的取值而产生接近于 0 的偏误是不清楚的，或者正的(或者负的)结果是否会主导总和产生一个较大的正的(或者负的)偏误也是不确定的。

作为一个对偏误的级别和符号的“不可预知性”的解释，当超过一个回归因子被排除的时候，我们可以比较方程 3.1 中两组变量被排除的偏误，其中一组是同时删掉 FAT 和 METABOLISM，另一组只去掉了 METABOLISM。表 5.2 提供了这种比较。当我们排除 FAT 和 METABOLISM 时，对于 AGE 的系数估计量的偏误更加显著。但对于所有系数而言，并不一定会出现这样的情况。比如，HEIGHT 的估计量的偏误就比较少，偏误的变化也没有明显的形式。EXERCISE 的系数估计量在大小上倾向于被高估，但现在却被低估了，SMOKER 的估计量实际上是被低估了，而现在显然是被高估了。

关于设定误差的讨论提供了一种研究策略，其中，对这种误差的含义的评估成为回归分析的整体的一部分。第一步是对理论的发展，并把它用回归方程中进行“翻译”。如果分析人员相信这种方程满足高斯-马尔科夫假设，它就可以被设想为是对估计模型的参考框架进行评估。如果研究人员非常幸运，就能够在样本中观测到参考框架模型中的所有变量，而且参考方程也足够“短小”，在给定的样本规模下，就可以避免严重的多重共线性，进而模型估计量可以继续不受设定误差的影响。但是，在更加典型的例子里，由于参考方程中一些变量的数据是不能使用的，因此模型就会“太长”，估计量就会出现排除误差。对于必须要出现设定误差的这种希望所表现出来的诱人的反应是，当同步构建理论和与之完美契合的回归模型时，通过考虑数据的可用性以及多重共线性可能出现的情况来跳过构建理论的阶段，或者至少在估计阶段“把它混到方程中”。但是，如果理论构建和对估计模型的规范这两个过程合二为一，评价被排除的变量的影响的可能性就会因为估计问题被平白无故地牺牲掉而变得必要。

恰当的研究策略在以下几种情况中被考虑到，其中估计回归排除了参考框架方程中的自变量：(1)当一个或多个特定的自变量在估计样本中不能被观测到时；(2)当所有自变量本来能被观测到，但是限于时间和资源而不能收集所有变量的数据时；(3)当估计完全参考框架模型会导致严重的多重共线性时。

当样本中特定的变量无法被观测到，从而导致不得不去掉某些变量时，如果来自其他样本的理论或者经验证据让我们有信心认为，被去掉的每个变量与

包含在内的变量有最微弱的联系，那么我们可以自信地说，这种排除不会使包括在内的变量的系数估计量产生实质上的误差。如果被包括在内的和被排除的自变量有很大的相关性，系数估计量可能会出现严重的偏误。唯一的问题是，我们是否可以或多或少地来判断偏误的方向？我们从上文中看到，如果只有一两个变量被排除，可能会导致关于偏误符号的合理推断。但是如果有很多变量都是不可被观测的，可以确定偏误方向的可能性就非常大。

如果这一问题缺乏资源呢？每个在参考方程中的数据都潜在地能够被测量，但是收集所有自变量的数据又是不切实际的。这里，对设定误差后果的认识提供了一个合理的标准去决定可以去掉哪个自变量。通常，当一位研究者被问及回归分析的研究目标时，他或她会说，是为了发展一个对因变量的、完整的或者无所不包的解释。照字面来看，这种目标近似于寻找“真实的模型”。对所有的经验研究来说，客观的情况是，这个目标是无法达到的。但是通常而言，当为了达到研究目的不得不需要更多的特性时，分析人员会说他们的研究关注的是某一个自变量，或者相关的一小撮自变量。因此，尽管参考框架模型可能包含无数的自变量，研究者主要还是对小范围内的、我们称之为核心变量的子集感兴趣。如果研究者能够获得无偏的、精确的核心变量的系数估计量，他或者她会感到很满意。这表明，如果时间和经费约束会限制包括在回归中的变量的数量，那么研究目标必须包括核心变量，再加上那些从参考框架模型中得来的自变量的子集。这样就能尽最大的努力，在数据收集有约束的情况下，去获得对于核心变量而言“好的”斜率估计量。这一点暗示了，包含在估计回归中的最重要的变量是那些在样本中与核心变量联系最紧密的变量。可以最安全地排除的变量是那些与每个核心变量仅有微弱关系的变量。当然，我们无法知道样本中这些变量相关性的取值，但是关于这些变量在其他样本中的关系的信息或者理论（或者两者兼有）很有可能在选择合理地去掉哪些变量的时候起到基础性的作用。

最后一个可能性是，如果估计样本中几乎没有足够的样本，而在参考模型中有很多充足的自变量，或者自变量之间有足够多的相关关系，那么参考方程会出现严重的多重共线性。因为多重共线性是一个由于样本中信息不充分而

产生的问题，所以唯一一个完全令人满意的解决方法是利用更多的数据。增加样本规模总是能够提高系数估计量的精度，只要附加的样本的自变量取值与样本中变量取值的均值不相同，而且这些样本不会在自变量之间增加相关性(Kmenta，1986:439—440)。

由于数据可用性的限制，如果有人有足够的自信来假设参考模型中某些参数的取值(基于理论或者从其他研究中得到的经验证据)，那么比起用数据来得出所有的参数，这些“知识”可以被用来得到更加有效的系数估计值。[25]如果已有的知识不能够被利用，严重的共线性会导致接受一些偏误去取得估计量的精确性成为一个吸引人的交换。然而，当多重共线性(并非没有能力观察变量)成为排除某些变量的理由时，我们则不必推断是否存在设定误差而可以通过引入敏感性分析来直接展开对排除变量后结果的分析。这种分析方法包括：(1)估计“完整的”参考模型的系数(该模型会得到无偏误的估计量以及比较大的标准误)以及各种被不精确地描述的“子模型”，其中每个子模型都包括核心变量，但是排除了各种不同的自变量的集合(因此会产生有偏误的估计量以及比较小的标准误)；(2)评估各种估计方法之间主变量的系数估计值的稳定性。举个例子，假设有理论估计 $\beta_1$ 是正的，并且不管是在“完整”的参考模型中，还是在各种错误描述的子模型中，对于 $\beta_1$ 的点估计一贯为正，其取值范围从最小的 $b_1^L$ 直到最大的 $b_1^H$，另外，由于 $X_1$ 增加一个单位而导致 $b_1^L$ 在因变量上的增加可以被认为有强烈的影响。在这些条件下，即使是基于完整模型估计得来的 $\beta_1$ 的置信区间都会包含 0，看上去拒绝 $\beta_1=0$ 的零假设而支持研究假设 $\beta_1>0$ 是恰当的。相反，如果各种不同的子模型得出的 $\beta_1$ 估计量的取值范围非常广，那么几乎不可能的结论是，假定数据都是可用的，我们却几乎无法获得理论中主变量对于个人的影响的精确估计量。

为了总结这一部分，我应该注意到，所有关于设定误差的讨论都与一个假设紧密相连，即存在一个明确的参考模型，那么估计模型就可以与参考模型进行比较，从而分析在经验研究中“错误的”模型。如果研究者有两个或者多个竞争性理论来解释一种现象，同时这一经验研究的目的在于判断哪种理论是真实的，那么我所提出的分析设定误差的效果的方法就毫无用处了。一种替代性的

策略是估计一个“嵌套的”模型。例如，有两个竞争模型，Ⅰ和Ⅱ，来解释因变量 $Y$，模型为：

$$Y_j = \delta_0 + \delta_1 X_{1j} + \delta_2 X_{2j} + \varepsilon_{1j} \qquad \text{（模型 Ⅰ）}$$

以及

$$Y_j = \mu_0 + \mu_3 X_{3j} + \mu_4 X_{4j} + \varepsilon_{2j} \qquad \text{（模型 Ⅱ）}$$

有人会从模型Ⅰ和模型Ⅱ的自变量中得到估计嵌套模型的系数：

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_4 X_{4j} + \varepsilon_j \qquad [5.10]$$

那么，如果对 $\beta_1$ 和 $\beta_2$ 的估计量在统计上是显著的，但 $\beta_3$ 和 $\beta_4$ 的估计量不显著，一个合理的推断是模型Ⅰ是合理的。同样，如果 $b_3$ 和 $b_4$ 显著，但 $b_1$ 和 $b_2$ 不显著，模型Ⅱ就被假设为正确。在这种方法中，一个潜在的缺陷是，4 个 $X$ 在本质上具有多重共线性。因为这种多重共线性很有可能导致方程 5.10 中所有的偏斜系数的估计量都不显著，即使其中有一个模型是正确的。[26]

## 均值的误差项为零的假设

在标准回归模型假设中，干扰项的均值为 0。准确地说，在以下方程中进行假设：

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + \varepsilon_j,\ E(\varepsilon_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = 0$$

如果这个假设被违背，那么 $E(\varepsilon_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = \mu_j$，其中干扰项不恒等于 0。那么，这就不是以下这种情况：

$$E(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = \alpha + \beta_1 X_{1j} + X_{2j} + \cdots + \beta_k X_{kj} \qquad [5.11]$$

而是，

$$E(Y_j \mid X_{ij}, X_{2j}, \cdots, X_{kj}) = \alpha + \beta_1 X_{1j} + X_{2j} + \cdots + \beta_k X_{kj} + \mu_j \qquad [5.12]$$

其中有两种可能性，$\mu_j$ 对观测值而言是恒定的，或者 $\mu_j$ 是变动的。后者会产生更严重的后果。

首先，考虑当 $\mu_j$ 为常数时，那么对所有观测值都可得到 $\mu_j = \mu$。这可以表示为，只要高斯-马尔科夫假设（除了本假设之外）能够被满足，偏斜系数的最小二乘估计也是无偏的。然而，OLS 截距估计量有 $\mu$ 个单位的偏误。为了确定 OLS 的截距估计量是否真的有偏，我们可以先让方程 5.12 中的 $\mu_j$ 取常数 $\mu$ 的取值，重新整理原式可得：

$$E(Y_j \mid X_{ij}, X_{2j}, \cdots, X_{kj}) = (\alpha + \mu) + \beta_1 X_{1j} + X_{2j} + \cdots + \beta_k X_{kj}$$

很明显，如果用 OLS 回归分析法，那么从以上方程中可以看到，截距的估计量为 $\alpha + \mu$，而不是 $\alpha$。也就是说，我们得到的估计量为 $\alpha$，同时有 $\mu$ 个单位的偏误。

对于因变量中的测量误差，我们可以理解为，每一个观测值的取值全都统一“扣除”了一个固定的数字，这样就制造出了一个不为 0 但其均值恒定的误差项。例如，假设样本中女性的体重已经按同一个“医生”的标准，在测量的时候全都少测了 5 磅。在这种情况下，估计方程就不再是方程 3.1 了，而是以下方程：

$$(\text{WEIGHT}_j - 5) = \alpha + \beta_C\ \text{CALORIES}_j + \beta_F\ \text{FAT}_j + \beta_E\ \text{EXERCISE}_j + \cdots + \varepsilon w_j$$

在方程两边都加上 5，就可以得出等价的方程，

$$\text{WEIGHT}_j = \alpha + \beta_C\ \text{CALORIES}_j + \beta_F\ \text{FAT}_j + \beta_E\ \text{EXERCISE}_j + \cdots + (5 + \varepsilon w_j)$$

其中，有一个均值为 5 的误差项。

那么当误差项的均值为 $\mu_j$，$\mu_j$ 随着观测值的变化而发生变化时，又会怎样呢？在标准回归模型中，满足所有高斯-马尔科夫假设，因变量的期望值完全被参数（$\alpha$，$\beta_1$，$\beta_2$，…）以及（如方程 5.11 的）自变量取值决定。但是，当干扰项的均值随着样本的变化而变化时，$Y$ 的期望值由回归参数 $X$ 的取值和 $\mu$（如方程 5.12）决定。实际上，$\mu$ 变成了一个被排除在回归方程之外的相关变量，从而导致了偏斜系数估计量的偏误。

总体而言，由排除法引起的设定误差会导致误差项的取值随着观测值的变化而不同。具体来说，当至少一个被排除的变量与至少一个包含在内的自变量相关时，这种情况才会发生。相反，当所有被排除的变量与每个包含在内的变量都不相关时，误差项的均值恒定且不为 0。[27]

当存在截断样本时，回归方程的干扰项也会出现不恒定的均值。也就是说，当回归分析被限制在某些有观测取值的样本中时，均值或高于某些固定值，或低于某些固定值。假设我们只考虑样本中那些体重不超过 150 磅的女性，为了简单起见，我们考察一个二元模型，总体而言，女性的体重由她的身高以及误差项决定[28]：

$$\text{WEIGHT}_j = \alpha + \beta\,\text{HEIGHT}_j + \varepsilon_j \qquad [5.13]$$

假设该方程满足高斯-马尔科夫假设。图 5.2 中的实线表示总体回归方程，黑点和叉号（共同）表示来源于女性的总体随机样本。其中黑点表示体重小于 150 磅的观测值，而叉号表示体重大于 150 磅的观测值。

对于体重小于 150 磅的女性，误差项的期望值与自变量 HEIGHT 负相关。为了解释其中的原因，首先须考虑，对于 WEIGHT 的期望值而言，HEIGHT 取值为 150，在图 5.2 中，这些值被标记为 $H^*$ 。而在限定总体中，当 HEIGHT = $H^*$ 时，误差项 $\varepsilon$ 必须为负或为 0。如果 $\varepsilon$ 为正，那么 WEIGHT 大于 150 磅。因此，在 $H^*$ 点，$E(\varepsilon_j \mid \text{HEIGHT}_j)$ 应该为负。其次，考虑到 HEIGHT 的一个取值 $H'$ 只是略微小于 $H^*$ 。因此取值为 $H'$，在限定女性群体中，$\varepsilon$ 必须为负或为 0，或是很小的正数，这样才能使 $E(\varepsilon_j \mid \text{HEIGHT}_j)$ 在 $H^*$ 上没有那么负向。确实，在身高的取值没有达到 $H^*$ 的时候，$\varepsilon$ 的最大正值等于图 5.2 中回归直线与水平线（WEIGHT = 150）之间的纵向距离。因为当 HEIGHT 减小时，$\varepsilon$ 的最大取值也随之增大，而 $E(\varepsilon_j \mid \text{HEIGHT}_j)$ 仍为负数，但是由于 HEIGHT 的减小，HEIGHT 的期望值开始接近于 0。因此，在那些体重小于 150 磅的女性中，$E(\varepsilon_j \mid \text{HEIGHT}_j)$ 与自变量 $\text{HEIGHT}_j$ 呈负向相关，且对 OLS 的估计量有 $\beta$ 个单位偏误。从图 5.2 中我们能清楚地看出，基于限定样本的估计回归直线会低估真实的斜率，因为最符合那些黑点（而不是叉号）的直线——在图中用虚线表

示——比总体直线更加平坦。

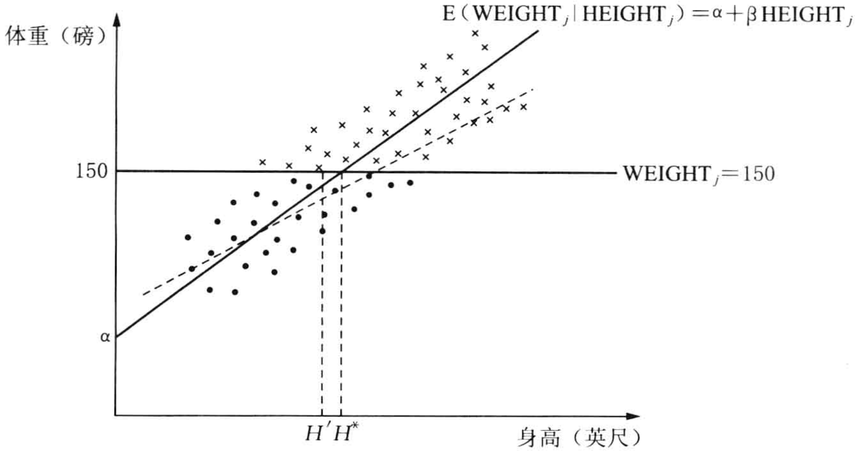



**图 5.2 "误差项的期望值为 0"这一假设的反例**

在体重样本中对斜率系数的低估是一种典型情况。在大多数实际应用中，如果在总体模型 $Y_j=\alpha+\beta X_j+\varepsilon_j$ 中用一些有限制的样本进行估计，这些样本里的观测值或来自总体中的样本 $Y_j$ 大于某个特定常数，或来自总体中的观测值 $Y_j$ 小于某个特定常数，斜率系数估计量会趋向于 0，从而产生偏误，即 $|E(b)|<|\beta|$。[29] 通过这样的形式来限制样本从而得到的估计量可以被认为是选择性偏误。选择性偏误的发生来源于总体的样本过少或过分地反映了一种或多种样本。在方程 5.13 的例子中，体重轻的人被过分看重了。还有一种类似选择性偏误的情况是描述教育对年收入的影响的二元变量回归分析。我们使用最常见的数据资源，利用类似限定样本的数据(即那些被雇用的人)来进行估计。由于我们排除了所有失业者，即那些没有收入的人，因此斜率系数的估计量会倾向于低估教育和收入在总体人口中的关系的强度。

当然，也有可能出现其他形式的选择性偏误。在前文的例子中我们可以看出，观测值被包含在估计样本中的可能性是由观测值在因变量上的取值决定的。另外，一个观测值被包括在样本中的可能性是由其他变量决定的。总体而言，当出现了选择性偏误时，斜率系数估计量就会出现偏误，除非认定加入一个

案例的变量在样本中与其他每个自变量都不相关。例如,我们假定方程 3.1(体重模型)的样本是通过在居民区的电线杆上张贴广告而招募来的志愿者。由于招募广告是按照街道号码来贴的,因此那些住在街道号码为奇数的街区的女性,比住在街道号码为偶数的街区的女性更容易看到招募广告。在这种情况下,一位女性被招募进入样本的可能性就是她家的地址是奇数还是偶数这一变量的方程。但是,我们有理由假设一个人的街道号码是奇数还是偶数与方程 3.1 中每个自变量没有关系。因此,选择性偏误可能不会在斜率系数估计量中产生偏误。相反,如果估计量的样本是自己选择的,年纪大的女性(或者有更多空闲时间的女性)可能比年轻女性更有可能成为志愿者而参与研究,这时斜率系数估计量就会出现偏误。

## 对于测量层次的假设

假设 A1 要求回归中的自变量必须量化或者二分,而且因变量是量化、连续和无限的。[30] 在回归模型中,观测值在因变量上的取值被假设与自变量、参数(即 $\alpha$ 和 $\beta_i$)以及误差项有关的方程。那么因变量必须可以取到任何由这一方程得出的数值型的取值。这就是为什么要求假设因变量是连续且无限的原因。[31] 严格来说,实际运用中没有变量是完全连续的,即使是一个物体的“长度”,只有当它可以被精确地测量时,才能被认为是连续的。否则,当它无法被严格精确地观测时(例如使用 1/10 英尺)[32],我们只能认为它是近似连续的。尽管没有严格的指导手册规定假设中的变量必须有多连续才能被视为达到“近似”的状态,一些变量——比如用美元来衡量的个人的净价值——显然是接近连续的。

对于这些不同种类的离散变量(即不连续的变量),其中只有少数适用于回归模型,而其他的则不适合。图 5.3 总结了不同的类型。离散变量可以是二分的或定性的[33],也就是说,有 3 个或更多无序的取值(例如,种族属性可以被归为 5 个种类:亚洲裔、非洲裔美国人、西班牙人、高加索人或者其他),或者有 3 个或更多有序的取值(例如,家庭中孩子的数量或者参加小型竞选的候选人的数

量)。[34]在有序的离散变量中,我们须把量化变量与非量化的变量区分开。

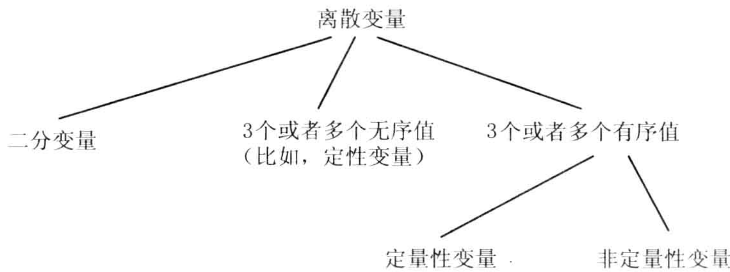



**图 5.3 离散变量的各种形式**

为了让次序离散变量作为因变量时能够更加适应模型,次序离散变量必须被量化。但是,我们并不能确定次序离散变量是否总是能用数字来表示。当然,即使用有次序的整数来标注这类变量,也不一定能够使它量化。例如,当次序离散变量有 3 种取值(低、中、高)时,有人用 1、2 和 3 来进行标注。但是只有当"低"和"中"的距离与"中"和"高"的距离相等时(其中"距离"指的是变量所代表的这一属性在数量上可被测得的差异),变量才能够量化。例如,如果宗教容忍度是这种属性,只有当高容忍度和中等容忍度的差异与中等和低容忍度之间的差异相同时,才能说变量是能够被量化的。

在特定的情况下,我们把次序离散变量当做连续变量也是可行的,这样更有利于我们把它当做因变量放在回归模型中展开分析。具体来说,当一个连续的次序离散变量有一个广泛的取值(比如一个组织中雇员的数量)时,把这个变量当做连续变量可能是合理的。相反,把任何只有少数几种取值(比如 5 个或者少于 5 个)的次序离散变量作为连续变量则是不恰当的。在这两种极端情况之间的选择就不那么明显,但是,作为一种总体的指导,量化的次序离散变量能取到的值越多,假设它是近似连续性也就越合理,把它当做回归方程中的因变量也就越合理。

作为离散回归量,定性的(即无序离散的)变量和非量化的次序离散变量无法被恰当地运用在回归模型中。尽管如此,通过运用两个或者多个二分变量,可以将这种类型的变量与因变量的影响合并,而这些二分变量可以用来表示样本是否符合特定的取值(对于如何正确构造二分回归量,参见前文;同见 Gujara-

*ti*, 1988:第 14 章;Johnson et al. , 1987:182—192; Schroeder, Sjoquist & Stephan, 1986:56—58)。[35]

对于回归模型,我们应该详细地考察二分因变量的结果。考虑到标准回归模型(方程 2. 2),其中 $Y$ 只能取 0 和 1。重新构造方程,使它能够把左边的误差项分离出来,我们能得到:

$$\varepsilon_j = Y_j - (\alpha + \sum_{i=1}^{k} \beta_i X_{ij})$$

当把 $\varepsilon_j$ 移到左边以后,我们可以看到,如果 $Y$ 只能取 0 和 1,那么对于 $X$ 的每个值,$\varepsilon_j$ 也可以只取两个值:$1-(\alpha+\sum_{i=1}^{k}\beta_i X_{ij})$ 和 $-(\alpha+\sum_{i=1}^{k}\beta_i X_{ij})$。也就是说,有了二分变量 $Y$,就会违背正态分布的误差项假设。误差项的变化也可以表示为如下形式:

$$VAR(\varepsilon \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = E(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) \cdot$$

$$[1 - E(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj})] = (\alpha + \sum_{i=1}^{k} \beta_i X_{ij}) \cdot [1 - (\alpha + \sum_{i=1}^{k} \beta_i X_{ij})]$$

这一方程澄清了误差项的变化会随着自变量的取值而系统地变化,这样也违反同方差假设(Aldrich & Nelson, 1984:13)。

但是当因变量是二分变量的时候,最严重的问题是系数可能出现"无意义"的解释。在因变量为 0 到 1 取值的例子中,$Y$ 的期望值必须相等(1 乘以 $Y$ 等于 1 的概率加上 0 乘以 $Y$ 等于 0 的概率),或者用以下方程来表示:

$$E(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = [1 \cdot P(Y_j = 1 \mid X_{1j}, X_{2j}, \cdots, X_{kj})] + [0 \cdot P(Y_j = 0 \mid X_{1j}, X_{2j}, \cdots, X_{kj})] \qquad [5.14]$$

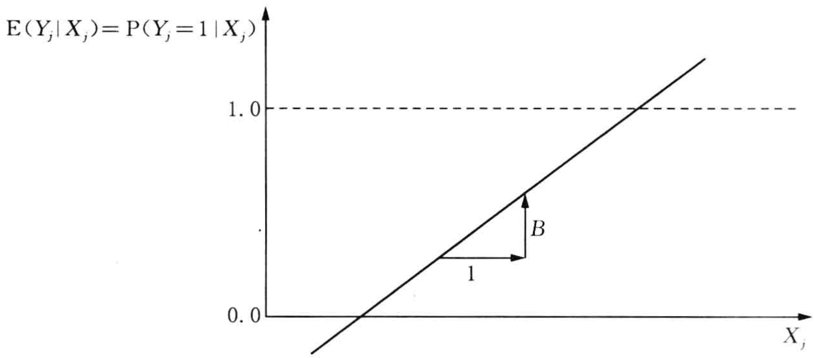
因为方程中最右边的一项总是为 0,方程 5. 14 可以化简为:

$$E(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = [1 \cdot P(Y_j = 1 \mid X_{1j}, X_{2j}, \cdots, X_{kj})]$$

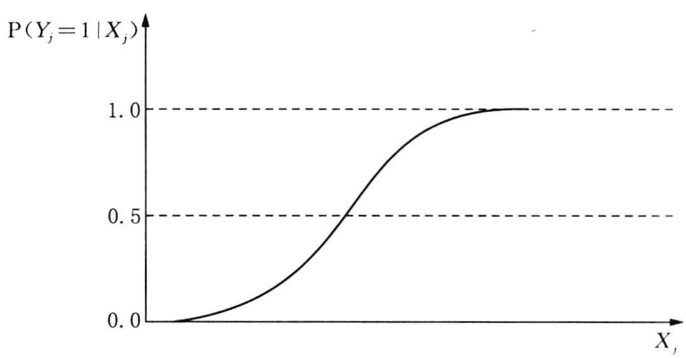
这说明,当因变量只能取 0 或者 1 时,$Y_j$ 的期望值可以被解释为 $Y_j$ 等于 1 的概率。[36]但是对于 $E(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = \alpha + \sum_{i=1}^{k}\beta_i X_{ij}$,并没有严格限制 $Y_j$ 取值范围在 0 到 1 之间。因此,$Y_j$ 等于 1 的概率可能会取无意义的取值,

比如小于 0 或者大于 1 的取值。

此外,在大多数因变量为二分变量的例子中,回归的线性假设并不可靠。图 5.4(a)中的二元变量模型也反映了线性假设以及因变量的期望值可以无限取值的特性。在这个例子中,回归模型的线性假设说明,即使当 $P(Y_j = 1 \mid X_j)$ 是 0.01 或者 0.99,自变量在概率上的影响也能通过斜率系数精确地反映出来。但是在绝大多数情况下,假设自变量的影响整体上减弱,正如 $Y$ 等于 1 的概率接近于 0.00 或者 1.00 会更合理。在这些情况下,logit 或者普罗比模型更加适合用来反映非线性的解释(参见 Aldrich & Nelson, 1984)。[37]后一种解释在图 5.4(b)中的二元变量例子中反映出来,利用 logit 模型得出的结果与该图在形状上非常相近。




(a) 对单个自变量 $X$ 的线性模型




(b) 对单个自变量 $X$ 的(非线性)普罗比模型

**图 5.4　因变量为二分变量时的二元模型**

## 无测量误差的假设

为了理解无误差测量回归假设的实质含义，我们有必要确定一个正式的测量模型。绝大部分的测量误差的形式可以用真实值（或者概念）$T$ 来构造；$T$ 可以用指数（或者观测值）$I$ 来测量；而 $j$ 指的是观测对象的取值；$f$ 指的是方程；$v$ 既指的是真实值也指误差项：

$$I_j = f(T_j, v_j) \qquad [5.15]$$

尽管我们都很希望能实现无误差测量，但是这需要我们把想要测量的概念在理论上弄清楚。例如，测量一个变量的真实取值是有可能的，比如说，“在最近的总统选举中，成年人参与州选举的比例”并没有可预见的错误。但是考察概念“生活的满意度”，则需要在调查中通过建构指标来测量。实际上，我们很难准确地测量这一变量的真实值，因为没有明确的规则来决定每个受访者认为能够反映“满意度”的项目，且这些项目能够组合成这一变量的精确的指标。

区分几种类型的测量误差是绝对必要的，其中每种类型都可能在实质上不同的情况下出现。第一种关键的区别在于随机和非随机误差。

### 随机测量误差

出现随机测量误差的关键是，误差项与真实值不相关（即 $COV[T_j, v_j] = 0$）。[38]“传递”数据中出现的误差最有可能出现随机误差，比如，在从文件的编码页中重新编码，或者从电脑文件夹的编码页中输入数据时出现错误。当调查采取封闭式的问题从受访者那里获得数据时，受访者任何单纯的猜测都会导致随机误差。同时，访问者的厌烦也会导致这样的误差。同样，模棱两可的问题可能会导致更大的随机误差。然而，也不能认为回应调查时出现的所有误差都是随机的，因为在回答中可能出现系统偏误（例如，那些不能投票表示他们受到社会压力的个人，以及人们在接受调查时，出现了抬高收入或者虚报年龄的趋势），这些都可能导致非随机误差的产生。

当受到因变量的限制时，随机测量误差（RME）对回归分析来说，麻烦是最小的。在这种情况下，总体估计量仍会保持无偏[39]，但是这些估计量的作用就不那么有效了，同时 $R^2$ 的取值也会变小（Berry & Feldman, 1985:28; Johnson et al., 1987:327—329）。当回归模型中的自变量出现 RME 时，参数估计量是有偏的，其偏误的程度的估计量是用测量误差的等级以及与自变量的关系的方程来确定的（Berry & Feldman, 1985:327—329）。只有在二元模型中，我们才比较容易判断源于一个自变量的 RME 的偏误的方向，其中单一斜率系数估计量的期望值总是比真实总体的取值小一些（即 $|E(b)|<|\beta|$）。[40]

## 非随机测量误差

任何不以单纯的随机形式出现的测量误差都被称为"非随机性"。与 RME 的情况不同的是，非随机测量误差（NRME）总是会在 OLS 估计量中造成偏误。但是这类偏误的特性以及它到底有多大的害处，还要根据误差的形式来判断。在非随机误差的种类中，Namboodiri 等人广泛地区分了作为被测量的变量的方程的误差以及由于"外生"变量导致的误差。即使是那些作为被测量的变量方程的误差，也包含很多不同的种类（Namboodiri et al., 1975:575）。在某些情况下，下列误差可以是线性的：

（1）截距误差：$I_j = T_j + \delta$，其中 $\delta$ 为常数，所以对于所有的观测值，指标值比真实值高估或者低估一个绝对的常数值。

（2）尺度误差：$I_j = \mu T_j$，其中 $\mu$ 为常数，所以指标值比真实值高估或者低估一个常数百分比。

（3）既有尺度误差也有截距误差：$I_j = \mu T_j + \delta$，其中 $\mu$ 和 $\delta$ 是常数。

但是当误差作为被测量变量的方程时（这种情况更加普遍），指标值可能与真实值有非线性的关系并有一系列不同的方程形式。

只有当研究者对回归方程中的截距（$\alpha$）有极大的兴趣时，才需要关注截距测量误差的持续性。因为偏斜系数估计量和 $R^2$ 值完全不会受到这种误差的影响。而且，如果一个截距误差的大小（即在定义模型中 $\delta$ 的取值）是已知的，OLS 截距估计量就可以通过测量误差造成的偏误来纠正。假设在估计方程（方

程2.2)中,用于估计的数据在自变量 $X_1$ 中被系统偏误高估了 $\delta$ 个单位。那么,尽管我们相信方程2.2是被估计的,但实际上是以下方程被估计了:

$$Y_j = \alpha + \beta_1(X_{1j} + \delta) + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + \varepsilon_j$$

我们把包含 $X_1$ 的项乘出来,可得:

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_1 \delta + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + \varepsilon_j$$

重新调整各项可得:

$$Y_j = (\alpha + \delta\beta_1) + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + \varepsilon_j$$

所以,要纠正本例中原始的截距估计量受到NRME的影响,我们可以简单地加上估计量 $\delta$ 乘以偏斜系数估计量 $\beta_1$ 得出的结果。当因变量受到截距误差影响时,也可以用类似的方法得到正确的系数。如果 $Y$(在方程2.2中)的指标值全部大于真实值 $\delta$ 个单位,截距估计量可以通过减去 $\delta$ 来纠正测量误差。

我们也可以确定,如果误差的大小已知,要更正尺度误差偏误所需的步骤。如果用因变量 $Y$ 代替真实值,那么如果我们测量的是 $\mu Y$(其中 $\mu$ 为常数),截距估计量和所有的偏斜系数估计量都会被夸大 $\mu$ 倍,那么就需要通过把所有的估计量除以 $\mu$ 来纠正。例如,如果我们通过调查了解到,在已知的总体中的女性把她们的体重比真实值报低了10%,那么所有的参数估计量都有向下的偏误,正确的取值需要将原始的取值除以0.90得到。[41] 相反,如果要得到一个自变量 $X_i$ 的真实值,而实际的记录量为 $\mu X_i$,只有OLS的估计量 $\beta_i$ 有偏误,那么偏误可以通过将估计量乘以 $\mu$ 来纠正。

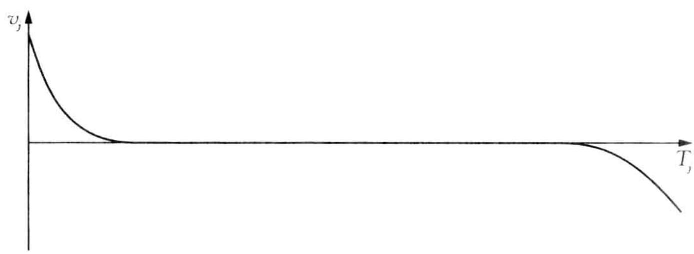
非线性误差(指作为被观测的变量方程的误差)的范围是没有界限的,因为任何不是截距或者尺度误差(或者两者结合的误差)的误差类型都是非线性的。例如,如果所有受调查者报低她们的体重,但是所报低的体重数(以百分比来计算)随着真实体重的增加而增加,那么体重较重的女性倾向于更大程度的“撒谎”,从而导致的NRME可能是非线性的。只有当以方程形式出现的非线性误差是显著的时候,才有可能把参数估计量更正过来。例如,如果报低的体重值符合方程 $\text{WEIGHT}'_j = \text{WEIGHT}_j - \mu\text{WEIGHT}_j^2$,其中 WEIGHT′ 指的是相对真实值

WEIGHT 而言，有误差的指标值，在这种情况下参数估计量的偏误就可以被更正。当然，如果要用最少的努力解决这类问题，就需要在进行回归分析之前就调整好体重指标值的取值，这样就能够在回归估计中使用没有误差的因变量。

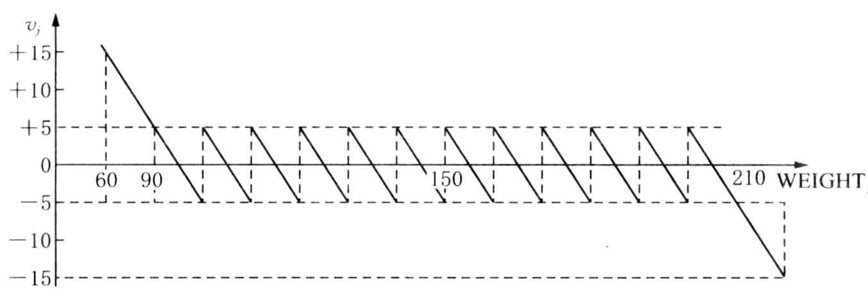
我们需要特别关注的是两种由被测量变量组成的关于非线性误差的特殊形式。一种可以被称为“趋于温和的误差”。受调查者在调查中回应某些问题并在某些社会压力的影响下表现出“温和”的一面时，会出现这种误差。Namboodiri 等人（1975：578）推测，当在调查中要测量“自由主义—保守主义”这类意识形态时，这两个方向的极端主义者会表现出比他们真实的观点更温和的态度。这种温和的表态所导致的 NRME 会出现类似于图 5.5(a)中的形式。如果图中的 $T_j$ 表示“自由主义的意识形态”，那么该图表明，意识形态的观测值反映了无误差的“温和的”真实值。但是指标值对“极端自由主义者”向下偏斜，对“极端保守主义者”向上偏斜。同样，如果体育锻炼变量在我们的体重案例中可以通过自我报告提供的指标值来测量；其中“宅女”会倾向于说自己会参加一些锻炼。然而“体育爱好者”倾向于报告适中的运动量，图 5.5(a)可能也能准确地描绘在真实值中的测量误差 EXERCISE。

对于作为被测量变量方程的 NRME，其另一种普遍的形式是由于分类导致的误差。[42]这种情况通常出现在连续变量被一系列有序类别的物体测量的时候。当一位分析人员使用已经被归类的二手数据时，这种误差是无法避免的。但是一些研究人员相信，分类能够在一定程度上减小误差，因此当他们认为这些变量的测量值都有误差的时候，他们在展开研究的时候会把连续变量分类。但是分类并不能一劳永逸地解决随机测量误差的问题，实际上也无法解决非随机误差的麻烦。总体而言，对于任何已经存在的测量误差而言，这一过程只是增加了 NRME 的一种新的形式。例如，假设 WEIGHT 是一个对真实体重的测量值，但是我们把它以 10 磅为间隔进行分类，及 100—110，110—120，…，190—200，并把样本中所有的女性归到这些类别的中点上(即分别归为 105，115，…，195)，同时将所有体重小于 100 磅的人记为 95 分，所有体重超过 200 磅的人记为 205 分。这种分类产生的测量误差反映在图 5.5(b)中，其中真实值和测量误差的大小之间的关系是一系列无连接的斜率为负的片断。那些体重恰好为 95 磅、105 磅或者 205 磅

的人，分类以后所得的观测值并不会产生误差，但是所有其他样本的测量值在分类以后都会有误差，对于体重在 90 磅到 210 磅之间的人，测量误差的范围总是小于 5 磅，但是对于非常瘦或者非常胖的人来说，体重的取值会反映出更大的误差。




(a) 趋向温和的误差




(b) 由于体重测量分类导致的误差

注：a. 这一图形描绘了在测量模型中 $T_j$ 和 $v_j$ 的关系，$I_j = T_j + v_j$，其中 $I$ 是真实值 $T$ 的指标。如果不存在测量误差(那么对所有的 $j$，$v_j = 0$)，描绘这种关系的曲线就会是那条水平线本身。

b. 这一图形描绘了测量模型中 $WEIGHT_j$ 和 $v_j$ 的关系，即 $WEIGHT'_j = WEIGHT_j + v_j$，其中 WEIGHT′是 WEIGHT 的指标。

**图 5.5 非线性的 NRME 作为被测量变量的方程模型的两种形式**

这种扭曲来源于由分类导致的测量误差，当类别数量减少时，这种扭曲会变得更加严重。尽管上文提到的 12 级体重类别(绝大部分是以 10 磅为间隔)可能能够相对减少系数误差导致的偏误，但实际上将体重取值分成两部分会造成更加严重的扭曲。在极端的例子中，由分类导致的测量偏误实际上可以改变参数估计量的期望值，使它从正数变为负数(或者相反)，因此导致结论出现严重

的实质性错误。举一个例子，如果我们在我们假设的总体中，用二分测量法测量体重，即 1(“胖的”)或者 0(“瘦的”)，并以 150 磅作为这两个类别之间的一个人为设定的界限。

表 5.3 中第一列表示的是方程 3.1 中的参数估计量的期望值。第二列表示的是这些参数估计量的期望值与真实的参数相比较的结果(即当 WEIGHT 被测量且没有误差时期望的参数值)。“有误差的”参数估计量的期望值与“没有误差的”估计量的期望值在两个自变量上正负号不同，即 EXERCISE 以及乘积SMOKER · EXERCISE——吸烟和锻炼对体重的影响的实质性解释。正如第 3 章中提到的，真实的参数表明，在不吸烟者中，从体育锻炼中消耗的能量每增加 100 卡路里/天，预计会导致 11.8 磅体重的减少(当所有其他的自变量都保持不变时)。然而在吸烟者当中，同样由于运动导致的能量消耗的增加，平均而言，只能造成 0.86 磅体重的减少。用二元(有偏误的)因变量进行分析，在不吸烟者中，保持其他所有自变量不变，每天增加 100 卡路里的消耗量，会增大 0.13 的变“胖”概率。但是在吸烟者中，增加同样强度的运动量只会导致提高 0.11 的变“胖”可能性。[43]

**表 5.3　基于分类范畴的测量误差的示例：体重变量[a]的“对分”产生的影响**

| 自变量 | 参数 | (1) 使用 NRME 对参数的期望值进行估计[b] | (2) 总体取值 (无测量误差)[c] |
|---|---|---|---|
| 截距 | $\alpha$ | −1.19 | 38.10 |
| CALORIES | $\beta_C$ | 0.00059 | 0.0291 |
| FAT | $\beta_F$ | −0.032 | −3.098 |
| EXERCISE | $\beta_E$ | 0.0013 | −0.1183 |
| HEIGHT | $\beta_H$ | 0.0102 | 1.346 |
| AGE | $\beta_A$ | −0.0038 | −0.285 |
| SMOKER | $\beta_S$ | 0.167 | 3.01 |
| $FAT^2$ | $\beta_{FF}$ | 0.00105 | 0.084 |
| SMOKER · EXERCISE | $\beta_{SE}$ | −0.0024 | 0.1097 |
| METABOLISM | $\beta_M$ | −0.289 | −1.795 |

注：a. WEIGHT 的测量为 0 = “瘦的”(< 150 磅)，1 = “胖的”(> 150 磅)。

b. 使用 OLS 回归对所有 134 个女性样本的体重进行回归，其中体重采用二分法标记。

c. 如正文所述。

关于外生变量方程的 NRME 也可以通过方程 5.15 规范化。在这里，$v$ 不是 $T$ 的方程，但它是其他变量的方程。在一些情况下，这种 NRME 的形式可以很容易被预测其结果。例如，因变量（$Y$）的指标值（$Y'$）的误差在回归模型中与其中一个自变量（$X_i$）相关，那么可以得到 $Y'_j = Y'_j + \mu X_{ij}$ 来代替方程 2.2 对 $Y'$ 的表达，再加上一个经过处理的项。这表示：(1) 对于所有参数，除了 $\beta_i$ 以外的估计量完全不受测量误差的影响；(2) $\beta_i$ 的有偏估计量可以通过减掉常数 $\mu$ 来更正。我们的例子中可能会出现这种形式的 NRME，其中 $Y$ 表明 WEIGHT，$X_i$ 代表 EXERCISE，并且 $\mu < 0$。如果体重是受调查者自己报告的，而且对"超重"这个问题，那些经常锻炼的女性比那些不经常锻炼的女性更加"敏感"，那么报低体重的幅度可能与体育锻炼量有直接关系。

但是对这种由外生变量构成的，有更加复杂的形式的模型，会导致结果更加难以预测。例如，受调查者的报告全部低于真实体重，那么，那些同等身高下比较胖的人更倾向于撒谎，测量误差的幅度就被假设为 $v_j = [(1.2)(\text{WEIGHT}_j/\text{HEIGHT}_j)]^2$。

**表 5.4 基于随机变量的测量误差的效果示例：体重的测量误差看做体重本身和身高[a]的非线性函数**

| 自变量 | 参数 | (1) 使用 NRME 对参数的期望值进行估计 | (2) 总体取值（无测量误差）[b] | (3) 测量误差导致的百分比偏差[c] |
|---|---|---|---|---|
| 截距 | $\alpha$ | 26.56 | 38.10 | −0.30 |
| CALORIES | $\beta_C$ | 0.0262 | 0.0291 | −0.10 |
| FAT | $\beta_F$ | −2.705 | −3.098 | −0.13 |
| EXERCISE | $\beta_E$ | −0.1085 | −0.1183 | −0.08 |
| HEIGHT | $\beta_H$ | 1.426 | 1.346 | 0.06 |
| AGE | $\beta_A$ | −0.255 | −0.285 | −0.11 |
| SMOKER | $\beta_S$ | 2.68 | 3.01 | −0.11 |
| $\text{FAT}^2$ | $\beta_{FF}$ | 0.073 | 0.084 | −0.13 |
| SMOKER · EXERCISE | $\beta_{SE}$ | 0.0996 | 0.1097 | −0.09 |
| METABOLISM | $\beta_M$ | −1.612 | −1.795 | −0.10 |

注：a. 这里真实的分数 $\text{WEIGHT}_j$ 是由 $\text{WEIGHT}_j - [(1.2)(\text{WEIGHT}_j/\text{HEIGHT}_j)]^2$ 来测量的。

b. 如正文所述。

c. (列(1)变量取值－列(2)变量取值)/列(2)变量取值。

在这种情况下，因变量中的测量误差是关于被测量变量以及其中一个自变量的非线性的方程。表 5.4 中的列(1)显示了当因变量被有误差的指标值替换时，方程 3.1 中系数估计量的期望值。这些参数估计量的期望值可以与列(2)中的真实的总体参数值相比较，其产生“相对”偏误的测量显示在列(3)中。尽管表5.4 显示，NRME 导致的偏误并不足以改变任何系数的符号，参数估计量还是高估了 HEIGHT 影响的大小，而其他变量的影响大约都被低估了 8%到 13%。

## 代理变量

最后，关于测量误差的警告是关于代理变量的。当一个变量不能被直接测量时，其他与这个变量的概念相关的变量就被假设来测量这个概念。例如，一个国家的人均收入通常被用来表示这个国家的发展水平(例如，Dye，1966；Pryor，1968)。在体重模型中，如果饱和脂肪摄入量的数据(FAT)对样本中的案例不适用，我们就可以考虑用女性吃快餐在所有用餐中所占的比重来作为替代变量。当使用代理变量后，即使对代理变量本身的测量没有一点误差，分析人员也可能会得出错误的结论。例如，没有任何误差的人均收入数据并不能证明当收入被当做代理变量时，对发展水平这一概念的测量没有误差。结果往往是，研究者必须警惕当代理变量被用做指标值时，可能出现的两种测量误差：(1)在对代理的真实值进行测量的时候，随机的或者非随机的误差；(2)由于代理变量的真实值无法有效地反映原变量而导致的非随机误差。

经济学家和其他社会科学家非常喜欢使用代理变量，特别是当他们用消费数据测量“需求”、“支持”或者“兴趣”时。例如：(1)通过测量家庭对一种产品的消费来计算他们对该商品的需求；(2)通过测量个人对某项原则相关的社团的贡献情况来估计他们对这项事业的支持；(3)通过统计某项体育比赛在社区内的售票情况来测量社区居民对该项运动的兴趣。[44]对于这些消费变量，我们必须对测量市场消费(或者售票情况)时产生的误差，以及由于产品或门票滞销带来的误差非常敏感，因为这些误差可能反映出人们对产品的精确的市场需求、支持程度或者感兴趣的程度。在那些作出贡献的人之中，人们贡献的大小可以测量对该项事业的支持程度，且误差非常小。但是，如果需要最小程度的付出

(不为0)去刺激贡献,那么,他们的行动与那些从来没有作出任何贡献的人相比,就会出现实质性的误差。由于支持该项事业被认为对其的投入一定不能小于0(也就是说,当其为0时,会被删除),因此,这些刚刚越过所谓的支持标准的人,同那些完全反对这项事业的人所获得的态度评分是一样的。这就会导致NRME,其中误差的大小是真实值的方程。对于那些真实得分低于标准线的人(被认为持支持的态度),支持的力度越小,那么其贡献被高估的部分就越大(实际上为0)。[45]

在一些情况下,当我们使用代理变量去测量一个概念时,这一变量可能就会从量化变为非量化,因此我们实际上测量的是另一个概念。当代理变量与被测量变量单调相关时——但不是线性相关——就会出现这种情况。一个家庭中,孩子的数量无疑是一个(离散的)量化变量,而不同等级在取值上的差异精确地反映了不同样本之间,子女数量的"差距"。然而,当子女数量被用做代理变量去测量"父母照顾子女所花费的时间"时,没有孩子和一个孩子的家庭所花费时间的差别,要比4个孩子和5个孩子之间的差别大得多,这一假设是非常合理的。这意味着,当我们用照顾子女的时间来测量子女的数量时,子女数量这一变量可能就不再是量化的了。当我们用代理变量去测量其他概念时,即使是连续变量也有可能失去其量化的特征。例如,Carter(1977)声称,收入与社会地位是非线性关系。因为以年收入1万美元为分界线,收入高于这一分界线的人群的社会地位与收入远比这一数字低的人群相比,其社会地位的差别要远大于他们与那些收入在更高水平的人之间的差异。结果,当收入被用于测量社会地位的代理变量时,这一指标就不再是量化的。如果一位研究者对概念和代理指标之间的非线性关系的特殊性质非常有信心,则可以用数学转换方式在恰当的测量等级范围下来"伸展"和"收缩"差距,并重新构建定量指标。因此,Carter(1971)推荐用收入的对数来衡量社会地位。

## 线性和可叠加性的假设

当判断一个线性可叠加模型是否适用于具体的研究应用时,我们需要考

虑，对于每个自变量而言，它与因变量的期望值之间关系的斜率是否依赖于这个自变量本身的取值以及/或者其他自变量的取值。如果理论表明，$Y$ 的期望值的变化是由自变量 $X_1$ 的微小的固定变化导致的（即理论认为 $X_1$ 和 $Y$ 的期望值之间关系的斜率）并依赖于 $X_1$ 的取值，那么就需要一种非线性的解释来说明这一问题。如果理论预测 $Y$ 的期望值的变化与 $X_1$ 的微小增量相关依赖于一个或者多个其他自变量的取值，那么称这一模型是非可叠加性的或交互性的会更恰当。

如果需要解释为什么一定要满足非线性和/或者非可叠加性，那么我们可以说，非可叠加性和/或者交互作用的特性会决定我们是否应该运用普通最小二乘假设来估计参数，或者需要抛弃 OLS 而采取其他的估计方法。如果一些数学的转换方法可以把非线性的和/或者非可叠加性的模型转换为对等的线性可叠加模型，这个模型就被称为内在的线性或可叠加。[46]同样，这个模型被称为是关于变量是非线性的和/或者非可叠加性的，但却是关于参数是线性的和/或者可叠加的。一个非线性的和/或者非可叠加的模型在经过恰当的转换以后，满足高斯-马尔科夫假设，其本质上是线性的和可叠加的，也可以利用 OLS 回归进行恰当地估计。相反，本质上不是线性的和/或者可叠加的模型——即关于参数不是线性的和/或者可叠加的——则不能利用 OLS 回归合理地进行估计作为替代，但可以用非线性最小二乘法或者极大似然估计法进行估计（Fox，1984：206—213；Greene，1990：335—340；Kmenta，1986：512—517）。

我们可以给出一个关于参数是线性的和可叠加的回归模型的例子：

$$f(Y_j) = g_0(\alpha) + \beta_1(X_{1j}, X_{2j}, \cdots, X_{rj}) + \beta_2 g_2(X_{1j}, X_{2j}, \cdots, X_{rj}) + \cdots + \beta_k g_k(X_{1j}, X_{2j}, \cdots, X_{rj}) + h(\varepsilon_j) \qquad [5.16]$$

其中，按照惯例，$\alpha$，$\beta_1$，$\beta_2$，…，$\beta_k$ 都是参数；$f$ 是关于变量 $Y$ 的方程；$g_0$ 是关于截距的方程；$h$ 是关于误差项的方程；每一个 $g_1$，$g_2$，…，$g_k$ 都是关于 $r$ 个一系列自变量的方程。通过对这些方程（即 $f$、各种不同的 $g$ 以及 $h$）进行数学变换，如取幂函数、取对数、对变量做乘法或者除法以及其他变换方式，方程5.16考虑了各种形式的、关于变量的非线性和/或者非可叠加性形式。

例如，方程 3.1 中的体重模型包含了一个运用了关于 FAT 的“平方”项，用来表述在脂肪摄入量和体重之间的非线性关系。[47]在另一个例子中，方程 3.1 中的一项(SMOKER・EXERCISE)是一个关于 SMOKER 和 EXERCISE 的“乘积”方程。对于这个乘积项的解释是，SMOKER 和 EXERCISE 对女性的体重有交互作用。对其他普遍的固有线性和可叠加模型的非线性和/或者非可叠加的形式，还有以下几种描述方式：

$Y_j = \alpha + \beta(1/X_j) + \varepsilon_j$ （双曲线或倒数模型）[48]

$\log Y_j = \alpha + \beta X_j + \varepsilon_j$ （半对数模型）

$\log Y_j = \alpha + \beta \log X_j + \log \varepsilon_j$ （指数模型）

$\log Y_j = \alpha + \beta_1 \log X_{1j} + \beta_2 \log X_{2j} + \log \varepsilon_j$ （对数模型）[49] [5.17]

其中，最后一种非线性非可叠加的模型被称为“科布-道格拉斯方程”，其最常见的表达方式如下[50]：

$$Y_j = \alpha X_{1j}^{\beta_1} X_{2j}^{\beta_2} \varepsilon_j \qquad [5.18]$$

通过在方程两边同时取对数，这一方程可以转换为方程 5.17 的形式。但是，我们将方程 5.18“线性化”为这种形式，取决于一个对 $Y$ 有倍增影响的误差项的模型。相反，以下方程则是一个与方程 5.18 类似的模型，但是有一个常规的可叠加干扰项：

$$Y_j = \alpha X_{1j}^{\beta_1} X_{2j}^{\beta_2} + \varepsilon_j$$

该方程的参数既非线性也不可叠加的，因为无论我们通过何种方法，都无法将其转换为对等的线性可叠加方程。

如果研究者在理论上指出了模型的线性和/或者交互性，但是在实践中却运用关于变量的线性可叠加估计模型(即模型中不包括关于变量的“转换了的”方程)，则会出现设定误差。在一些情况下，不规范的方程形式可以对应地被解释为由于排除相关变量而导致的设定误差。例如，考虑方程 3.1 中的体重模型的参考框架以及包括在参考模型中的所有自变量的估计模型，在一个线性可叠加模型中忽略了两组相关变量，即吸烟与运动对体重的交互作用以及在脂肪摄

入量与体重之间的非线性关系：

$$\begin{aligned}\text{WEIGHT}_j = {} & \alpha + \beta_C \text{CALORIES}_j + \beta_F \text{FAT}_j + \beta_E \text{EXERCISE}_j \\ & + \beta_H \text{HEIGHT}_j + \beta_A \text{AGE}_j + \beta_S \text{SMOKER}_j \\ & + \beta_M \text{METABOLISM}_j + u_{Wj} \qquad \text{（估计模型）}\end{aligned} \qquad [5.19]$$

在这种情况下，参考模型的自变量的偏斜系数(即 CALORIES、HEIGHT、AGE 以及 METABOLISM)与因变量有线性且可叠加的关系。通过分析，由于从回归模型中排除回归量 $\text{FAT}^2$ 以及(SMOKER · EXERCISE)而导致的偏误，我们可以看到由这种错误的方程形式导致的扭曲。假设我们已知由于排除变量导致的设定误差，就可以得出如下结论：参数估计量对任何与 $\text{FAT}^2$ 和 SMOKER · EXERCISE 相关的自变量都会产生偏误。

但是对于与因变量有非线性关系的自变量 FAT，或者对于与因变量有交互影响的变量 EXERCISE 和 SMOKER，方程 5.19 中对 OLS 估计量的扭曲远超过普通意义上的“偏误”。这是因为在估计模型中，参数被估计——不考虑它的期望值——有一种与参考模型假设本质上不一致的解释。线性可叠加方程 5.19 假设，所有自变量对 WEIGHT 的影响都是恒定的，它们不会因为自变量取值的不同而产生变化。结果，如果回归模型 3.1 中的参数被错误地用方程 5.19 进行估计，任何假定恒定斜率的估计量本质上都会错误地被表述。因为 FAT、SMOKER 以及 EXERCISE 在不同自变量的取值上产生的影响的强度是有差异的。

在这种非常特殊的情况下，我们就可知，错误地将关于变量非线性且不可叠加的模型当做线性且可叠加的模型会有何种后果。例如，假设参考模型是一个二阶多项式：

$$Y_j = \alpha_j + \beta_1 X_j + \beta_2 X_j^2 + \varepsilon_j \qquad \text{（参考框架）} \qquad [5.20]$$

其中，自变量 $X$ 由其取值与均值的偏差来衡量。[51] 假设研究者错误地认为 $X$ 与 $Y$ 的关系是线性的：

$$Y_j = \alpha_j + \beta_1 X_j + u_j \qquad \text{（其中 } u_j = \varepsilon_j + \beta_2 X_j^2\text{）} \qquad [5.21]$$

我们可以看到，OLS 估计量对 $\beta_1$ 期望值可以通过以下方程给出（Theil，1971：550）[52]：

$$E(b_1) = \beta_1 + \beta_2[E(X^3)/E(X^2)]$$

因此，由于设定误差导致的偏误为 $\beta_2[E(X^3)/E(X^2)]$。$E(X^3)/E(X^2)$ 的符号是由自变量分布[53]的偏态决定的。如果 $X$ 的分布是对称的，那么 $E(X^3)$ 等于 0[54]，因此偏误也为 0。如果分布是倾斜的，那么偏误的符号是由 $E(X^3)$ 以及 $\beta_2$ 所决定，因为 $E(X^2)$ 总是为正。当 $X$ 的分布是正向倾斜时（$X$ 有非常大的取值，因此 $X$ 的均值也比中位数要大），$E(X^3) > 0$，但是当分布是负向倾斜时（那么中位数会比均值大），$E(X^3) < 0$。因此，我们能够得出结论说，当 $X$ 由其偏离于均值的程度来衡量，同时 $X$ 的分布是对称之时，OLS 估计尽管没有正确地描述方程 5.21 中的线性模型，但是仍然能得出无偏估计量 $\beta_1$。同样，方程 3.2可以表示对自变量 $X$ 的任意给定值 $X^*$，$X_j$ 与 $E(Y|X_j)$之间关系的斜率在方程 5.20 中等于 $\beta_1 + (2\beta_2 X^2)$。因此当 $X = 0$ 时，其斜率就为 $\beta_1$。这意味着，当 $X$ 由其对自身均值的偏离程度来衡量且对称时，一般来说，对这个错误描述的线性模型的估计，将会在 $X$ 的均值处产生一个斜率系数估计量，且等于多项式曲线（方程 5.20）的斜率。这看上去是一个令人满意的结果。例如，考虑到多项式模型，其中表示 $X$ 和 $Y$ 的期望值关系的斜率会在一个最低点 $c$ 和最高点 $d$ 之间变动。而这两点是当 $X$ 变化时，所能取到的最大值和最小值。合理的解释是，当这种模型被错误地认为是线性模型的时候，该模型倾向于生成一个恒定的、大于 $c$ 且小于 $d$ 的斜率。

这种性质也描述了方程 3.1 中的多元体重模型。对于这一模型，FAT 和 WEIGHT 期望值之间关系的斜率在 0.26 和 5.30 之间变化。因为 FAT 从其在总体中取值的最低点（20 克）到最高点（50 克）之间变化（参见图 3.1）。[55] 表 5.5 显示了方程 3.1 中由于假设该模型为线性的而得出的错误效果以及排除了 $FAT^2$ 后的结果。表格的列(2)确定了 $\beta_F$ 估计量的期望值在线性模型中取值的范围是 0.26 到 5.30，当取值为 1.381 时，真实的总体取值又根本性地偏离了 −3.098。正如表 5.1 所示，FAT 和 $FAT^2$ 有强烈的相关性（达到了 0.98）。因

此，从模型中删除 $FAT^2$ 必然会导致 $\beta_F$ 出现偏误。但是我在表 5.5 中没有描述对 $\beta_F$ 估计量的偏误的测量，原因是这个偏误——其范围是从 1.381 到 -3.098——实际上是没有意义的，因为这是用不正确的模型进行的不准确的估计。在没有详细说明的线性模型中，$\beta_F$ 反映了脂肪消耗量对体重作用的强度被假设为当消耗量变化时，体重保持稳定。在非线性模型中，脂肪消耗量对体重作用的强度被假定随着消耗量的变化而变化。$\beta_F$ 表示了当脂肪消耗量为 0（消耗量的一个阶段，实际上，恰巧在所反映的总体取值区间之外）[56]时，其对体重的影响。在其他消耗量的水平上，这种效果的强度是 $\beta_F$ 和 $\beta_{FF}$ 的方程。

**表 5.5 指定函数形式错误影响的示例：将方程 3.1 中的非线性/非可叠加性模型看做现行/可叠加性模型进行处理**

| 变量 | 参数 | (1) 总体取值[a] | (2) 错误指定为线性模型且不考虑 $FAT^2$ 的情况下对参数的期望估计 | (3) 错误指定为可叠加性模型且错误规定误差导致的百分比偏差[b] | (4) 不考虑 SMOKER · EXERCISE 的情况下对参数的期望估计 | (5) 错误规定误差导致的百分比偏差[c] |
|---|---|---|---|---|---|---|
| 截距 | $\alpha$ | 38.10 | 33.98 | -1.89 | 28.06 | -0.26 |
| CALORIES | $\beta_C$ | 0.0291 | 0.0307 | 0.05 | 0.0295 | 0.01 |
| FAT | $\beta_F$ | -3.098 | 1.381 | ** | -2.984 | -0.04 |
| EXERCISE | $\beta_E$ | -0.1183 | -0.0748 | -0.37 | -0.0939 | ** |
| HEIGHT | $\beta_H$ | 1.346 | 1.426 | 0.06 | 1.447 | 0.08 |
| AGE | $\beta_A$ | -0.285 | -0.136 | -0.52 | -0.267 | -0.06 |
| SMOKER | $\beta_S$ | 3.01 | 5.08 | 0.69 | 4.91 | ** |
| $FAT^2$ | $\beta_{FF}$ | 0.084 | — | — | 0.082 | -0.02 |
| SMOKER · EXERCISE | $\beta_{SE}$ | 0.1097 | 0.0739 | -0.33 | — | — |
| METABOLISM | $\beta_M$ | -1.795 | -1.750 | -0.03 | 0.01 | 0.01 |

注：a. 如正文所述。

b. （列(2)变量值－列(1)变量值）/列(1)变量值。

c. （列(4)变量值－列(1)变量值）/列(1)变量值。

** 偏差的测量对估计量的意义并不大。

表 5.5 中的列(2)和列(3)明确说明了脂肪摄入量和体重的线性关系，但也会对其他体重的预测值的参数造成偏误。例如，当 $FAT^2$ 被排除在模型外时，在

那些不参加体育锻炼的女性中，吸烟对体重的影响（如 β 所反映的）倾向于被高估 69%。同时，$\beta_A$（反映年龄对体重的影响）的期望参数估计量可能会被低估 52%。

表 5.5 也反映了由于排除了乘积项 SMOKER · EXERCISE（参见列(4)和列(5)），因而错误地把“体重”参考模型当做可叠加模型后得到的分析结果。有人可能会预见，$\beta_E$在错误的可叠加模型中的期望值为 −0.0939，这一数值表示了 EXERCISE 和 WEIGHT 期望值关系的恒定斜率，处于非吸烟者的真实斜率($\beta_E = -0.1183$)和吸烟者的真实斜率($\beta_E + \beta_{SE} = -0.0086$)之间。但是无论是参加体育锻炼还是吸烟，对于女性体重的恒定影响的估计量在可叠加模型中永远都无法准确地反映真实的关系，即吸烟对体重的影响依赖于女性的运动量，并且运动带来的效果对吸烟的女性和不吸烟的女性还不相同。

对于表 5.1 相关关系的检视可以让我们预见到，相对于排除 $FAT^2$，排除 SMOKER · EXERCISE 可能导致的对其他变量的偏斜系数估计量的偏误并不严重，因为 SMOKER · EXERCISE 与所有变量之间的相关性（除了 SMOKER 和 EXERCISE 以外）都非常微弱。确实，对于所有偏斜系数（除了 $\beta_S$ 和 $\beta_E$），由错误的模型所得出的估计量的期望值总是高于或低于将近 8%（参见表 5.5 中的列(5)）。但是，一般没有理由指望错误的可叠加的交互模型会比被当做线性模型的非线性模型产生更少的偏误。

在很多情况下，我们选择线性和可叠加的模型来进行分析，是因为社会科学家没有弄清楚能否用理论来预测非线性或者不可叠加的关系。但是用“头脑风暴”的方法去思考可能影响因变量的自变量显然是不够的，因此在标准线性可叠加回归模型中加入变量的做法并不常见。如果我们相信模型是关于变量的线性的和可叠加的，这表明我们实际上作出了本质的假设：每个自变量对因变量的作用完全独立于自变量被固定时的取值。所以，对于模型规范展开的头脑风暴必须包括一套分析，即每个自变量对因变量的作用是可变的——随着自身或者其他自变量取值的变化而变化。[57] 只有当这些变化着的影响预期的性质被彻底弄清楚，才能选择一套合适的回归模型，从而展开有意义的分析研究。

## 同方差和缺乏自相关假设

同方差性假设(A6),即在回归模型中,误差项的条件方差保持恒定。在方程 2.2 中,$\mathrm{VAR}(\varepsilon_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = \sigma^2$,其中 $\sigma^2$ 为常数。当误差项的条件方差不为常数时,则会出现异方差性。那么用符号来表示则为 $\mathrm{VAR}(\varepsilon_j \mid X_{1j}, X_{2j}, \cdots, X_{kj}) = \sigma_j^2$。任意两个观测值的误差项都不相关的假设(A7)被称为“缺乏自相关”,或者“序列相关”。我们将分别对异方差性假设和缺乏自相关假设展开深入讨论,但也会简单地讨论违反了这两条假设的解释的后果,因为违反这两种假设的结果是一样的。正如我们所见,由于存在异方差性和自相关,OLS 系数估计量是无偏的,但并不能说满足 BLUE。一种替代的方法是广义最小二乘法(GLS),用这种方法可以得到满足 BLUE 的估计量。

### 自相关的本质含义

在回归方程中,对于所有关于误差项的假设,理解缺乏自相关性的实质含义必须要先搞清楚误差项代表了哪些变量对因变量产生的联合影响(但这并没有从回归方程的回归量中体现出来)以及因变量会出现的任意随机因素的影响。如果我们把这些被忽略的变量用 $Z_1, \cdots, Z_m$(见方程 5.3)来表示,我们就可以看出,缺乏自相关的假设要求对任意一对观测值($j$ 和 $h$)、被排除的变量以及对 $Y$ 的取值的随机因素的净影响—— $\delta_0 + (\sum_{i=1}^{m} \delta_i Z_{ij}) + R_j$ 和 $\delta_0 + (\sum_{i=1}^{m} \delta_i Z_{ih}) + R_h$ —— 是不相关的。

自相关作用特别类似在时间序列回归中所出现的问题。为了理解出现这种问题的原因,我们考虑在时间序列模型中,每个被排除的变量($Z_1, \cdots, Z_m$)组成的误差项是正向的自相关。也就是说,那些变量现在的取值与之前的取值是正相关的。[58]那些随着时间的推移而“递增”的变量会倾向自相关。具体来说,有很多倾向于随着时间的推移,整体上逐步增加的社会的、政治的和经济的变量(例如,加利福尼亚州的人口数量、个人的收入,或者某个组织或者政府的

花费)都是自相关的,那些保持稳定的主观态度也是自相关的。

现在,我们来考虑两个连续的观测值的误差项(在时间 $t$ 和 $t+1$ 上):

$$\varepsilon_t = \delta_0 + \delta_1 Z_{1t} + \delta_2 Z_{2t} + \cdots + \delta_m Z_{mt} + R_t \quad [5.22]$$

以及

$$\varepsilon_{t+1} = \delta_0 + \delta_1 Z_{1,\ t+1} + \delta_2 Z_{2,\ t+1} + \cdots + \delta_m Z_{m,\ t+1} + R_{t+1} \quad [5.23]$$

对于时间序列方程,我们将用 $t$ 代替 $j$ 来表示观测值,用来提示我们这个模型是随着时间推移而变化的。为了简化这些数学方程,我们假设所有的 $Z$ 都被调整过了,其均值都为 0。把这个均值代入方程 5.22 两边,可得[59]:

$$E(\varepsilon_t) = E(\delta_0) + \delta_1 E(Z_{1t}) + \delta_2 E(Z_{2t}) + \cdots + \delta_m E(Z_{mt}) + R_t \quad [5.24]$$

$E(\varepsilon_t)=0$(A4 中已经假定这一方程成立);通过假设,$E(Z_{1t}) = E(Z_{2t}) = \cdots = E(Z_{mt}) = 0$,并且 $E(R_t)=0$,因为 $R$ 代表 $Y$ 的内在的随机性。在方程5.24中,0 代替所有的均值,这样得到 $E(\delta_0)$ 也等于 0。因此,常数 $\delta_0$ 也必然等于 0。方程 5.22 和方程 5.23 可以化简为:

$$\varepsilon_t = R_t + \sum_{i=1}^{m} \delta_i Z_{it}$$

$$\varepsilon_{t+1} = R_{t+1} + \sum_{i=1}^{m} \delta_i Z_{i,\ t+1}$$

接着,我们利用协方差的定义可得[60]:

$$COV(\varepsilon_t,\ \varepsilon_{t+1}) = E(\varepsilon_t,\ \varepsilon_{t+1}) = E\left[\left(R_t + \sum_{i=1}^{m} \delta_i Z_{it}\right)\left(R_{t+1} + \sum_{i=1}^{m} \delta_i Z_{i,\ t+1}\right)\right] \quad [5.25]$$

在方程 5.25 中的表达方式可以改写为以下表达方式,对连续误差项的协方差:

$$\begin{aligned} COV(\varepsilon_t,\ \varepsilon_{t+1}) = {} & E\left(R_t + \sum_{i=1}^{m} \delta_i Z_{i,\ t+1}\right) + E\left(R_{t+1} + \sum_{i=1}^{m} \delta_i Z_{it}\right) \\ & + E(R_t,\ R_{t+1}) + \sum_{i=1}^{m} \sum_{j=1}^{m} \delta_i \delta_j E(Z_{it} Z_{j,\ t+1}) \end{aligned} \quad [5.26]$$

$R_t$ 和 $R_{t+1}$ 的均值为 0 表示，在方程 5.26 中，右边的前三项是在一个随机变量（$R_t$ 或者 $R_{t+1}$）以及其他变量之间的一个协方差，因此必须为 0（见注[60]）。因此，从方程右边"去掉"几项，只留下：

$$\mathrm{COV}(\varepsilon_t,\ \varepsilon_{t+1}) = \sum_{i=1}^{m}\sum_{j=1}^{m}\delta_i\delta_j \mathrm{E}(Z_{it}Z_{j,\ t+1}) \qquad [5.27]$$

那么，假设这个协方差为 0——正如回归假设所要求的那样——合理吗？

我们首先考虑方程 5.27 中，当 $i = j$ 时，子项的总和为：

$$\sum_{i=1}^{m}\delta_i^2 \mathrm{E}(Z_{it},\ Z_{i,\ t+1}) \qquad [5.28]$$

这个表达式包括了所有被排除的 $Z$ 经过加权的总和的协方差，其取值在连续的时间点上的取值之间。首先，所有的加权（$\delta_i^2$）的总和（即系数的平方）都是正的。另外，我们假设每个 $Z$ 都正向自相关，这意味着每个协方差[$\mathrm{E}(Z_{it}Z_{i,\ t+1})$]的总和都为正。因此，所有的项在加权的总和中都是正的，因此和本身也是正的。

但是我们并不能就此推导出在方程 5.27 中，当 $i \neq j$ 时所有项的总和的正负。但是，我们至少可以得出，因变量和一堆自变量的总体趋势都随着时间的推移而增加。我们可以预计，绝大部分的[$\mathrm{E}(Z_{it}Z_{j,\ t+1})$]项的总和都是正的，正如绝大多数 $Z$ 变量在时间点 $t$ 上与绝大多数 $Z$ 变量在时间点 $t+1$ 上正相关。同时，在这种情况下，大量的 $\delta$ 系数，即测量被排除的 $Z$ 对因变量的影响，也应该是正的。如果是这样，绝大多数 $i \neq j$ 时的项数的总和在方程 5.27 中都应该是正的。那么，当这些项被加入方程 5.28（其中 $i = j$）后，最终的误差项的协方差在连续观测值上的取值也应该是正的。这说明了在时间序列模型中，自相关性是一个普遍问题。

为了理解自相关作用的根本含义，我们举一个例子。请考虑在时间序列背景下的体重模型，其中我们假设用一个回归模型预测人们在每个星期开始时的体重。[61]在时间序列的条件下，包含在误差项中的一个最重要的变量可能是个人的健康。假设个人的健康与模型中的自变量不相关[62]，那么把健康从模型中排除，则不会引起估计量的偏误。但是个人的健康可能是自相关的。因为一个人在这

个星期的健康状况是预测这个人下个星期健康状况的非常好的指标。同时，个人健康的特征会对回归模型中的干扰项有正向的自回归作用。

有时候，那些产生自相关作用的被排除的变量可以被设想为某些对未来时期的系统产生“震撼”效果的大事件。例如，我们用时间序列模型来解释旧金山的平均房价，那么1989年的地震可能会导致房价在后来很长一段时间内的剧烈波动。如果事实如此，同时这件事情在模型中并没有被明确地提出，那么就会出现自相关作用。实际上，在这一模型中，真正被忽略的并不是地震本身，这一隐含变量实际上是“对地震的恐惧”，这种恐惧心理在1989年以后的若干年间，一直停留在较高的水平上。

如果受自相关性影响而“被排除”的变量的误差项也能够制造同样自相关的干扰项，那么时间序列模型中的设定误差所导致的自相关性就不值得惊讶了。为了理解为什么这一结论是正确的，假设误差项为 $\varepsilon$ 的参考模型被一个排除了自变量 $X_1$ 的回归模型估计。如果 $X_1$ 是自相关的，那么估计模型中的干扰项 $u_t = \varepsilon_t + \beta_1 X_{1t}$ 也有可能是自相关的。然而，在用排除法的设定误差的例子当中，当无法判断被排除的变量与自变量是否相关时，自相关性通常是一个次要的问题。这是因为那些由于设定误差而导致的系数估计量的偏误一般比自相关作用的后果严重得多。

在时间序列模型中，另一种可能导致自相关作用的设定误差的形式是一种不正确的方程形式。比如，假设参考模型的框架是非线性的多项式：

$$Y_t = \alpha + \beta_1 X_t + \beta_2 X_t^2 + \varepsilon_t \qquad \text{（参考框架）}$$

在图5.6中可得该曲线，但是线性模型

$$Y_t = \alpha + \beta_1 X_t + u_t \quad \text{（其中 } u_t = \varepsilon_t + \beta_2 X_t^2\text{）} \qquad \text{（估计模型）}$$

代替该多项式被估计。我们从图5.6中可以清楚地看到，不正确的方程形式会产生误差项 $u$，即它倾向于在自变量取较大值或较小值时取正值（即，如果 $X_t < w$ 或者 $X_t > z$，那么 $E(u_t \mid X_t) > 0$），而在 $X$ 的“中间”取值上为负（即，如果 $w < X_t < z$，那么 $E(u_t \mid X_t) < 0$）。如果 $X$ 是那种会随着时间的推移取值不断递增的变量，那么连续观测值的误差项也会相关。

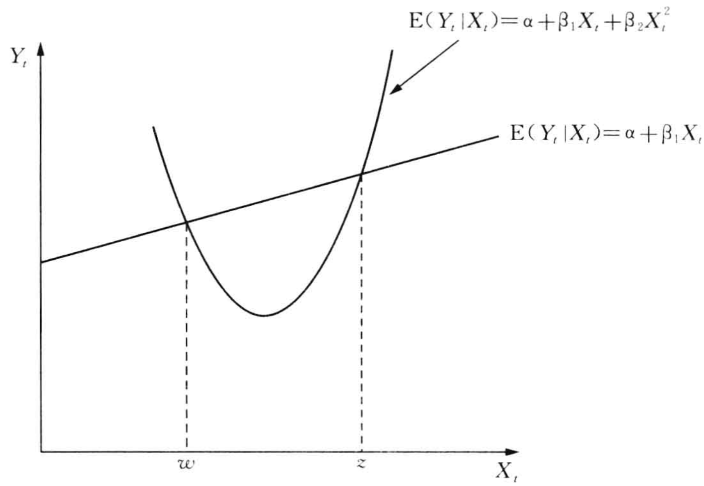



**图 5.6　在时间序列模型中，自相关与错误的函数形式的联系**

与时间序列模型相比，自相关性在横截面模型中总的来说并不是一个大问题，但是空间自相关可能在某些情况下出现。确实，一旦观测对象彼此之间出现了某种程度的"被建构"的关系，那么我们就应该怀疑出现自相关性的可能。在时间序列模型中就有这种情况，其中观测值在一个时间的序列中被构建。而当观测对象为独立的个人的时候，他们之间几乎没有或者根本没有任何联系(正如在调查研究中对一个国家的人口的随机抽样一样)，在这种情况下，对观测对象进行建构几乎是不可能的，因此自相关性也不是什么问题。但是在关于体重的案例中，如果我们研究的女性住在四人间的宿舍里，她们每个人都被这个小社区所建构，因此她们会花更多的时间与其他人在一起。我们再假设，女性的健康情况会影响她们的体重，但是却被排斥在回归模型之外。在共同的居住环境中，健康很有可能有空间上的相关性。因为传播疾病会在整个环境中蔓延，女性的健康可能会与整幢楼中邻居的健康状况相关。这样，自相关作用就又会发生。

另一种在分析中需要特别注意的空间自相关作用可能出现的情况是，被观测的单位是政治管辖区，比如大都市地区的城市、美国的州或者其他国家(Odland，1988)。在这种情况下，被排除在回归模型之外的用以解释辖区内社

会的、政治的或经济特征的变量常常有地区相关性。最后,假设我们希望用回归分析来检验哪些因素决定了同性恋神职人员的态度。[63]因为在总体人口中找到这些同性恋神职人员是件非常困难的事情,所以我们只能构建一个随机抽样来进行经验研究。事实上,我们可以利用滚雪球的方法。在雪球样本中,只有很少一部分人会被选中做访谈,但是每个人都会被要求从总体人口中确定其他同性恋神职人员(Sundmsn, 1976:210—211)。假设这些人有可能推荐他们的朋友和熟人,那么样本中的观测对象会被这种观测方式所"建构",包含在回归误差项中的一个观测对象在某些变量上的取值,就很有可能与其他观测对象的取值相关。

### 同方差性的实质含义

关于同方差性的一个重要的看法是,虽然它通常被看做关于一系列残差平方的假设,但同时也可以看做一系列关于因变量的方差的假设。如果 $k$ 个自变量——$X_{1j}$, $X_{2j}$, …, $X_{kj}$——$\mathrm{VAR}(\varepsilon_j \mid X_{1j}, X_{2j}, \cdots, X_{kj})$是定值(在某一个 $\sigma^2$ 下),那么在方程 2.2 中的误差项就可以被称为同方差性。但是对于 $X$ 的任意单一的一组值,$\alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj}$ 是定值,并且我们记 $\mathrm{E}(Y_j \mid X_{1j}, X_{2j}, \cdots, X_{kj})$ 为 $s_j$。关于这一组确定的随机变量的 $X$,从方程 2.2 中可以得出 $Y_j = s_j + \varepsilon_j$。因此,对于任意固定的一组自变量,$Y$ 和 $\varepsilon$ 仅仅相差一个常数。正是因为这一点,$Y$ 和 $\varepsilon$ 的条件方差必然是相等的。这就证明了我们的确可以将同方差性看做对于每一组自变量,条件方差 $Y$ 等于常数 $\sigma^2$ 这一假设。理解了同方差性这一巨大的意义,它就可以被用来考虑误差项和因变量的方差。

尽管自相关通常与时间序列有紧密的联系,但异方差性才是横截面研究中的主要问题。在一些例子中,异方差是从度量因变量的方差中导出的。具体而言,如果全体的测量误差在代表性研究的观测值之间有系统变化,那么就有了异方差性。比如,在我们之前所举的例子中,体重是由试验组中的女性自己提交的,并且所有的女性都试图更诚实,因此任何测量错误都是由于误解而不是刻意扭曲造成的。再比如,随着年龄的增加,她们对体重会更敏感,从而会更频

繁地测量自己的体重。在这种情况下,年长的妇女会比年轻的妇女提供更准确的数据。因此,对于不同组的自变量,即使她们体重的真实方差是常数,但是误差方差却会随着年龄的增加而递减,从而导致因变量的年龄和指标的方差之间的负相关。

另一种可能会因为测量误差而导致异方差性的情况是跨国研究。在这类研究中,国家在发展水平上的差异很显著。如果关于因变量的数据是从政府工作报告中得来的,那么这些记录的质量很可能随着政府的进步而提高(因变量的方差将会降低)。结果是,在回归方程中,一旦因变量和政府的发展水平呈现相关性(这并非不可能),这些变量也会和因变量的方差相关,从而导致异方差性。

上述例子表明,在一个独特的回归模型的应用中,评估模型是否具有异方差性时,最主要的问题是误差项的条件方差(也可以考虑自变量的方差)是否在模型中有可能和其中的一个或者多个自变量相关。如果相关,就可以预测存在异方差性。进一步而言,那些被认为与误差项的方差有联系的特殊变量的预测在处理异方差性的问题上,对模型的再设定以及估计过程非常关键。

如果异方差性的出现仅仅是由于因变量测量的随机误差而导致的,我们就可以用现有的工具来检查和处理。检查的方法可以是直接观测,或者更正规一点,我们可以调查 OLS 残差项和其他一个或多个自变量的方差之间的关系(Kennedy, 1985:97—98)。[64]另一方面,这种检查方式需要研究人员作出自变量与误差项的方差有关系的假设。当异方差性导致一个单独的自变量和 $\varepsilon$ 的方差相关时,可以在可视化的图表中画出在估计样本上的回归残差和可疑变量之间的关系,从而验证是否存在异方差(Berry & Feldman, 1985:78—80; Gujarati, 1988:327—329; Rao & Miller, 1971:116—121)。此外,还可以借助其他的检查技术,包括 Goldfield-Quandt 检验和 Glejser 检验(Berry & Feldman, 1985:79—81; Gujarati, 1988:329—336; Johnson et al., 1987:303—304),这些方法适用于仅有一个自变量的模型,更普遍的 Breusch-Pagan 检验(Johnson et al., 1987:304—305; Kmenta, 1986:294—295)则更加适用于误差项的方差和两个或者多个变量的线性组合相关的情况。一旦在一个样本中证实异方差

性检验，并且分析人员充分了解它的形式（也就是说，确切地知道到底是哪一个或者哪些自变量以何种形式依赖于变量 $\varepsilon$），就可以用广义最小二乘法（下面会对这一技术进行简单描述）得到一个完美的估计，从而就能解决异方差的问题。

但是，如果异方差并不仅仅是由于测量的质量随着观测值变化而导致的，那么这种情况下最好不要采用 GLS 来解这一问题。相反，异方差可以看做模型有序地再设定的一种表征。具体而言，异方差性可能是由“内部”自变量和“外部”自变量的交互作用导致的。在这种情况下，当“内部”变量与一个或者多个“已经包含在内部”的自变量相互作用的时候，解决异方差性的方法应该是重新划归一个或者多个变量到误差项中。

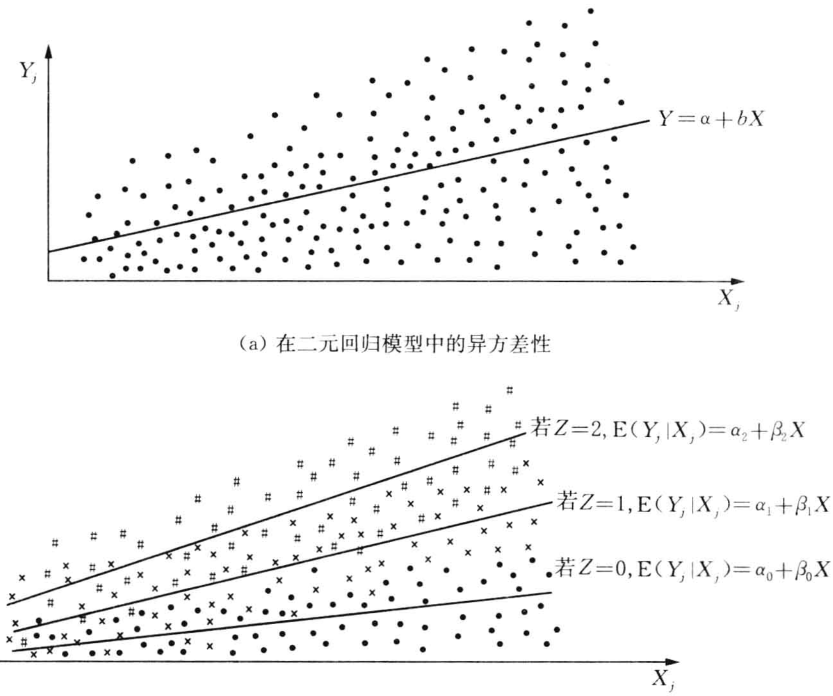
举个例子，我们介绍一个双变量的回归模型，该模型的分析单位是美国家庭，其中因变量是一年内家庭度假的开支（记为 VACATION），自变量是年收入（记为 INCOME）：

$$\text{VACATION}_j = \alpha + \beta\,\text{INCOME}_j + \varepsilon_j \qquad [5.29]$$

我们可以非常肯定地假设斜率 $\beta$ 是正数，这表明，随着家庭年收入的增加，家庭用于度假的支出的期望也会相应增加。但不仅是度假的平均支出会随着收入增加而增加，度假支出的方差（给定一个收入水平）也会随着收入的增加而增加。换言之，这个回归模型可能是具有异方差性的。这一论断逻辑上意味着，对于低收入家庭，度假的支出水平会比较低，那么其方差也会比较小，因为低收入家庭必须首先把家庭收入投入到家庭必需的方面，而只有较少一部分收入可以用来旅游和娱乐。但是随着家庭收入的增加，家庭可支配收入增多，家庭用于度假的支出的水平和方差都会增加。因此，我们关于收入和度假开支之间关系的假设就是，高收入是高水平度假开支的一个必要不充分条件。

事实上，我们在任何时候都可以认为，在一个观测数据中取值比较大的自变量都是导致因变量比较大的充分不必要条件，在这种情况下，就应该怀疑是否存在异方差性。这种“必要不充分”条件导致了观测数据的散点图呈现三角形。图 5.7(a)描述了观测值的两个变量 $X$、$Y$ 之间的散点图，其中 $X$ 取比较大的值是 $Y$ 取比较大的值的必要不充分条件。异方差可能是由于变量 $X$ 和其他

没有考虑到的但却影响 $Y$ 的变量 $Z$ 之间的相互作用因素造成的。一种可能性如图 5.7(b)所示，其中被排除的变量 $Z$ 是一个可以以 0、1 和 2 为取值的变量。图 5.7(b)和图 5.7(a)都是关于同一组观测样本的，但是从图 5.7(b)所示的模型中可以得到，对于 $Z$ 有任意固定取值的模型，$X$ 和 $E(Y_j \mid X_j)$之间呈现线性关系，而且误差项是具有同方差性的，但是 $X$ 和 $Y$ 的期望值之间的斜率却和 $Z$ 的取值相关。如果图 5.7(b)中所示的模型用 $Y$ 和 $X$ 的二元回归来描述，那么根据经验，这一错误的方法必然导致图 5.7(a)中所示的异方差的残差项。




(a) 在二元回归模型中的异方差性

(b) 由于被忽略的变量 $Z$ 导致的设定误差

注：# 表示 $Z=2$ 的观测值，× 表示 $Z=1$ 的观测值，· 表示 $Z=0$ 的观测值。

**图 5.7 “必要但不充分”关系的散点图**

回到前述的家庭年度假支出模型，很重要的一点是应该注意到异方差性质的产生并不仅仅是测量误差引起的。即使对所有家庭的年度度假支出的测量都非常完美，异方差性还是可能存在。我们或者数据分析人员面临的主要问题

是，在方程 5.29 中，哪些不在考虑范围内的变量被当做了误差项？一个还是多个这类变量与家庭收入存在交互作用，从而影响年度度假支出？“家庭的满意度（幸福感）产生于家庭度假的消费量”这一变量很可能是有问题的。一个更加合理的假设是，家庭用于度假的支出不仅受制于家庭年收入，而且也和度假能够给家庭带来的满意度（幸福感）相关。家庭收入和由度假带来的满意度交互作用，从而影响到度假支出。对于度假中的满意度（幸福感）很低的家庭而言，家庭收入可能对度假支出影响甚微，因为这类家庭只会花费小部分收入去度假而不理会年收入的总量。但是随着度假带给家庭的满意度（幸福感）的增加，年收入对度假消费的影响也会增大：那些感觉到度假能够给家庭带来很大满意度（幸福感）的家庭，如果年收入很高，那么他们会花费相当可观的一部分收入去度假，而年收入比较低的家庭可能仅仅能支出一小部分到度假上面。因此，在年度度假消费模型中解决异方差性的最好方法并不是 GLS 估计，而是应该用交互模型，这一模型理论上具有同方差性，

$$\text{VACATION}_j = \alpha + \beta_I \text{INCOME}_j + \beta_S \text{SATISFY}_j + \beta_{IS}[(\text{INCOME}_j)(\text{SATISFY}_j)] + \varepsilon_j$$

其中，SATISFY 表示从度假中得到的满意度（幸福感）的总量。这一修正后的模型回避了对异方差性的“技术手段”解法，而有利于对度假支出这一模型本质上的解释。[65] 因此，如果新模型误差项具有同方差性，那么 OLS 方法将会是一个恰当的估计手段。

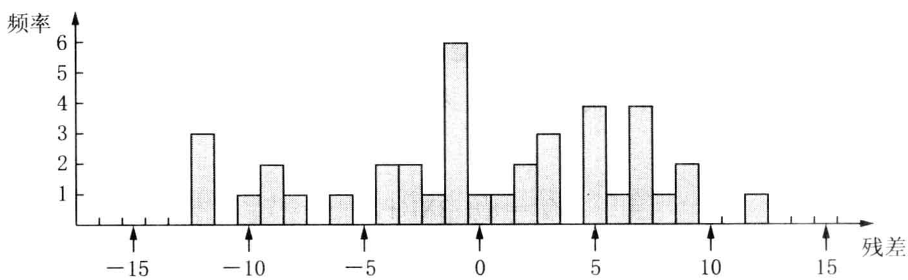
最后一个例子是体重模型。假设我们已经错误地设定了模型，漏掉了 EXERCISE这一变量，那么理论上会得到如下的可叠加模型：

$$\text{WEIGHT}_j = \alpha + \beta_C \text{CALORIES}_j + \beta_F \text{FAT}_j + \beta_H \text{HEIGHT}_j + \beta_A \text{AGE}_j + \beta_S \text{SMOKER}_j + \beta_M \text{METABOLISM}_j + \beta_{FF} \text{FAT}_j^2 + u_j \quad [5.30]$$

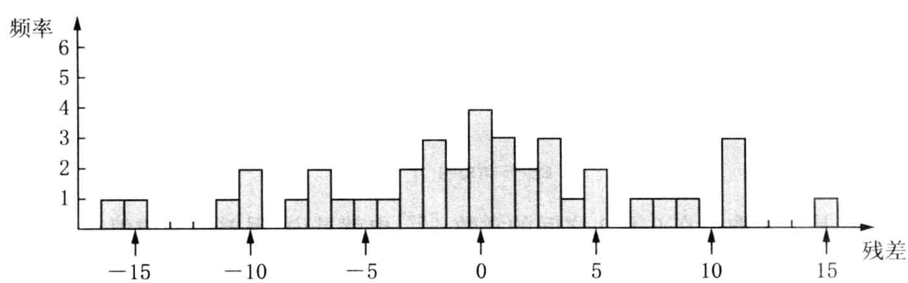
假设在这一人群中，SMOKER 的确和 EXERCISE 有交互作用，从而身体锻炼对体重的影响在吸烟者中的影响比不吸烟者的大。方程 5.30 可以用异方差来表征，从而对于不吸烟者，其误差项的方差大于吸烟者。假设在不吸烟者中，身体锻炼对于体重的影响很大，如果从模型中删除 EXERCISE，则会导致在预测有

大量身体锻炼和少量身体锻炼的人的体重时产生巨大的误差，从而使误差项的方差比较大。但是如果假定在吸烟者中，身体锻炼对于体重的影响比较小，从而删除 EXERCISE 这一项在预测这类群体的体重时，也不会导致太大的误差，使误差项的方差比较小。

关于这一异方差的预测，可以通过在收集到的关于女性的数据上对方程5.30进行GLS回归后，比较吸烟者和不吸烟者的残差的分布从而得到验证。图5.8是关于这两个分布的直方图。在图5.8(a)中，从人群中随机抽取了40个吸烟者；而在图5.8(b)中，从人群中随机抽取了40个不吸烟者（这样可以使两个直方图具有可比性，保持两个样本的容量相同是非常重要的）。通过直接观察图5.8中的图形，我们可以确信不吸烟者的残差分布比吸烟者的更加分散，这一点和我们关于异方差性的分析是一致的。




(a) 随机抽取40个吸烟者




(b) 随机抽取40个不吸烟者

**图5.8 方程5.30中OLS残差的分布**

## 异方差性和自相关作用的后果

正如前文所述，尽管存在异方差性和自相关作用，OLS 系数的估计量仍然是无偏的。这一结论凭直觉来看是有道理的。例如，在一个存在异方差性的二分回归模型中，

$$Y_j = \alpha + \beta X_j + \varepsilon_j$$

其中，$\beta > 0$，当其他标准回归假设都满足该条件的时候，误差项 $\varepsilon$ 的方差随着自变量 $X$ 的增加而增大。这意味着，当 $X$ 取值较大时，相对于较小的取值而言，其所对应的 $Y$ 更有可能偏离真实的回归曲线。在任意一个样本中，少量的观测值在 $X$ 取较大值且 $\varepsilon$ 取非常大的正值时，会使得 OLS 斜率系数的估计量大于 $\beta$。同样，在任何样本中，少量的观测值在 $X$ 取较大值且 $\varepsilon$ 取了一个非常大的负值时，则会使得 OLS 斜率系数的估计量小于 $\beta$。但是因为误差项的均值为 0(假设 A4)，$\varepsilon$ 取数值很大的正值和负值的可能性相等，所以如果随机样本量可以取无限大，那么平均斜率估计量仍然等于 $\beta$。举一个自相关性的例子，请看以下这个二分的时间序列方程：

$$Y_t = \alpha + \beta X_t + \varepsilon_t \qquad [5.31]$$

这是对同一个人在不同时间点的观测值。我们假定，除了缺乏自相关性以外，其他回归假设都被满足。因此，自相关性就以社会科学研究中最普遍的形式呈现出来：正向一阶自回归。在这种自相关作用的形式中，误差项在任意时间点的取值 $t$ 都可以用其本身在前一个时间点的取值 $t-1$ 以及随机变量 $u$ 组成的方程来表示[66]：

$$\varepsilon_t = \pi\varepsilon_{t-1} + u_t$$

其中，$\pi$ 是一个取值在 0 和 1 之间的常数。[67]也就是说，误差的期望值在一个阶段是前一个阶段误差的一个固定的比例。而且，误差项在取值上随着时间推移而变化的过程包含了两部分：其一是系统的成分(用参数 $\pi$ 来表示)，另一部分是随机成分 $u$。

假设方程 5.31 中的 $X$ 随着时间的推移会不断增长。图 5.9 以实线表示了

$X$ 与 $Y$ 的总体间的关系。那么，假设用一组样本数据来估计这个方程，那么第一次观测时（在时间点 $t'$）恰好是负值（如图 5.9 所示）。[68] 由于存在正向一阶自回归的误差结构，因此误差项很可能在下一个时间点也是负的。确实，当任意一个观测值都存在取值较大的负向误差项时，误差项很可能在多个观测点上都保持为负数。一种类似的表述是，取值较大的正误差项（比如图 5.9 中的时间点 $t''$）[69] 很可能导致其他的正误差项在多个观察点上也为正。因此，反映在图5.9 中的误差形式在给定的样本中是正常的。基于 OLS 回归模型的 $\beta$ 估计量，在图 5.9 中明显地高估斜率系数。但是在不断重复的样本中，估计量误差的均值会超过 0，因此一阶自回归误差结构在长期来看很可能同样会产生正的 $\varepsilon$ 和负的 $\varepsilon$。最终的结果是，OLS 对 $\beta$ 的估计量保持无偏。

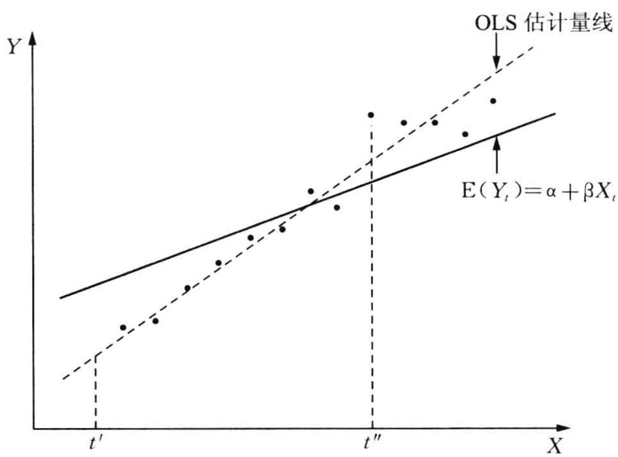



**图 5.9　自相关性的意义：正向一阶自回归过程**

但是对于存在异方差性或自相关问题的模型，OLS 估计量不再是 BLUE 的。取而代之的是，用广义最小二乘法得出 BLUE 估计量（见 Hanushek & Jackson，1977；Wonnacott & Wonnacott，1979）。更有效率的 GLS 估计是由最小的、加权的残差平方决定的（不同于 OLS 中未加权的残差之和）。观测值被认为有最大取值的误差项（已知异方差性或自相关作用的信息）被赋予最小的权重在总和中实行最小化。例如，当异方差性存在时，观测值的误差项被认为

会有一个比较大的取值，因为对于有较大方差的误差项而言，其在一个特定阶段的加权是很小的。[70]最后，在异方差性和自相关性的持续作用下，通常用来估计系数估计量标准误的方程是不正确的，它同时会对OLS估计量的标准偏移产生有偏的估计量。那么，照惯例计算出来的OLS估计量的置信区间和$t$检验都不再是合情合理的(Berry & Feldman，1985:77—78；Gujarati，1988:325—326；Hanushek & Jackson，1977:146；Johnson，1987:124—125)。[71]

## 误差项为非正态分布时的假设

我们已经看到，几种重要的OLS系数估计特征并不一定要求满足误差项是正态分布的这一假设。具体而言，只要高斯-马尔科夫定理仍然得到满足，那么就能保证OLS系数估计量保持无偏和有效。正态假设的原则性的重要意义是用来检验系数估计量的统计显著性以及构建置信区间的。当回归系数估计量基于一个小规模样本得出时，正态假设就会被要求来证明统计检验。在一个小规模样本中，正是误差项的正态分布假设让人们能够确定系数估计量的样本分布也是正态的。然而，统计学家已经说明，根据中心极限定理，当估计量由一个大样本得出时，即使方程的误差不是正态的，回归系数估计量的样本分布也是正态的。结果推导回归方程时，随着样本量的增加，人们对方程是否满足正态假设就不再那么关注了。

在回归方程中，误差项在回归模型中是正态分布的假设的最普遍证明还是依赖于中心极限定理。我们已经解释过误差项反映了大量影响因变量但是被排除在回归方程之外的自变量以及随机变量的影响(见方程5.3)。中心极限定理(及其扩展)表明(只有很少的例外)，当一系列独立的随机变量的总和趋近于无限时，这些变量的总和的分布接近正态(Greene，1990:109；Gujarati，1988:90；Hanushek & Jackson，1977:335)。因此，对于接近无限大的独立随机变量的总和的误差项的范围，我们可以证明误差项正态分布。然而实际上，我们很难反驳被排除的变量组成的回归误差项是独立的这一论点。幸运的是，有一些检验可以使误差项偏离正态性。Fox(1984:174—175)推荐了一种用目测回归残差的图标的方法来探测较大的正态性偏离，同时也可以用零假设的统计检验来验证残差的分布是不是正态的。[72]

# 第6章 | 结论

尽管绝大多数社会科学家可以把回归假设的各种正统定义倒背如流,但是他们很多人对这些假设的实质还没有足够的认识。除非理解这些假设的意义,否则回归分析几乎不可避免地成为一种死板的训练,其中有许多自变量被毫无根据地加到标准线性可叠加回归中,从而得出系数的估计量。尽管这种联系可能偶尔会得出可信的结论,但是这也只能看研究者的运气了。运用回归分析来获得本质上有规律可循的可信结论,则需要使用者注意两点:(1)是否每个回归假设都满足身边的每个具体研究项目,尤其有时候一些假设并不与项目相符;(2)违反这些假设应当如何解释。撰写本书就是为了鼓励学生不要把回归假设当做必背的一长串词汇,而应该是把它们当做一系列关键的且必须经过仔细分析才能运用的回归分析的条件。

# 注释

[1] 当然，经验研究被用来测试假设时，都必须有广泛的假设来进行支持，无论那项研究是实验性的还是非实验性的，或者依赖于定量方法（例如，普罗比模型、logit 模型或者二阶最小二乘法）或定性方法（例如，比较案例研究）。唯一要注意的是，展开分析的时候是否清楚地了解这些假设。

[2] 对回归分析的简介可以参考的文章有 Lewis-Beck(1980) 以及 Berry & Feldman(1985)。在计量经济学领域中可以找到更多全面介绍回归分析的书，比较好的、适中的读本是 Kelejian & Oates (1989)、Gujarati(1988)、Johnson, Johnson & Buse(1987)、Wonnacott & Wonnacott(1979) 以及 Hanushek & Jackson(1977)；更高级的还有 Greene(1990)、Kmenta(1986) 以及 Judge(1985)。

[3] 在本书中，当一个变量有两个下标时，第一个下标通常是表示在一系列变量中的一个单独的变量，而这一系列变量中的每一个都用同一个符号来表示（例如，$X$）。但是当一个变量用一个特殊的符号来表示时，则只需要一个下标来标记不同的观测对象或者变量假设中的特殊取值。那么，即使我用 $k$ 个回归量模型来表示自变量，即 $X_1, X_2, \cdots, X_k$，我仍会用 $X$ 在一个二元回归模型中标记唯一一个自变量。

[4] 相反，在回归模型中持续不断的误差项以及后面将看到的假设，都要求回归模型被假设是由 $Y$ 对固定取值的自变量的条件均值误差决定的。

[5] 为了简化讨论，我们假设在真实模型中的解释变量对因变量有线性和可加性。

[6] 一个二分（或者虚拟）变量是只有两个可能取值的变量。一个定量的（或者定距的）变量指的是：(1)分配给观测对象的数字，以便观测对象根据他们所拥有的财产来排序；(2)对于成对出现的对象，他们之间得分的不同精确地反映了他们之间所拥有的财产的差异。量化的变量既可以是连续的，也可以是离散的。连续变量指的是那些能够取任何数字的变量。相反，离散变量被认为是只能取特定的、有限的值。

[7] 注意，任意两个变量 $V_1$ 和 $V_2$，$COV(V_{1j}, V_{2j}) = 0$，当且仅当 $\rho_{12} = 0$（其中 $\rho_{12}$ 表示 $V_1$ 和 $V_2$ 之间的相关性）时是成立的，因为 $\rho_{12} = COV(V_{1j}, V_{2j})/\sigma_1\sigma_2$，$\sigma_1$ 和 $\sigma_2$ 分别表示 $V_1$ 和 $V_2$ 的标准差。如果研究者能够重复对自变量进行固定的抽样取值，那么就没必要引入假设 A5 了，因为当自变量的取值固定时，这一假设一定会满足（对固定回归量的回归模型的发展——没有假设 A5——参见 Wonnacott & Wonnacott, 1979）。但是社会科学家想要固定自变量的情况是非常少见的（可能会在实验的情况下出现），他们经常必须接受任何他们所能观察到的取值。因此（用专业术语来说），我们创造了回归模型，同时使自变量可以随机取值。

[8] 当把两个正态假设合并时(A8)，A7 指的是 $\varepsilon_j$ 和 $\varepsilon_h$ 是独立的。

[9] 最大似然估计法是替代估计模型的一个例子。幸运的是，如果满足高斯-马尔科夫假设，最大似然估计量与 OLS 的估计量是一致的。

[10] Isabelle Romieu 和 Walter Willett 提供了可以构建总体方程的数据。从我的分析中没有办法得到体重的决定因素的确定结论，Romieu 和 Willett 的数据（见 Romieu et al., 1988）已经被人为地“处理”来进行回归假设的解释。

[11] 在 134 名女性的总体中，严格说来，因变量在方程中是离散的，即违背了假设 A1。但是，这 134 个离散变量已经非常接近于连续变量了。

[12] 在 Romieu 等人(1988)的数据中，没有包含 METABOLISM 的数据。METABOLISM 只定义了对一个人而言，在正常生活状况下的典型的能量输出情况（Garrow, 1974）。但是鉴于已经有人人为地建立了代谢率的指标，那么我也任意地创建属于我自己的单位：运气。

[13] 这是因为对任意一个自变量 $X$，在回归方程中，$X$ 和因变量在任意 $X$、$X^*$ 期望值上的关系的斜率，指的是 $X$ 在 $X^*$ 方程衍生的程度。

[14] 对回归模型中交互项作用的拓展讨论，参考 Jaccard, Turrisi & Wan(1990)。

[15] 为了达到技术上的正确性，这些 OLS 估计量的性质只有当假设 A5 在每个 $X_i$ 都被满足的时候，$X_i$ 的分布才独立于 $s$；如果 $X_i$ 很少与 $s$ 不相关，只有当样本较大时，这些性质才能被满足(参见 Gujarati, 1988:57)。两个变量被定义为独立的，即一个观测值在一个变量的取值能够完全不受它在另一个变量的取值的影响。如果这两个变量之间有某些非线性的关系，那么它们可能是不相关的，但不能说它们是相互独立的。

[16] 无偏性是一个估计量所谓的小样本性质。这也就是说，无论样本的规模有多大，小样本性质都会得到满足。在一些样本中，有偏的估计量是一致的。一致的估计量指的是，当样本量趋于无穷大时，估计量的偏误和方差都趋于 0。

[17] 因此，在图 4.1 中的每个概率分布都可以转换为关于 $\theta$ 的抽样分布。关于抽样的含义，参见 Mohr(1990)。

[18] 在另一些情况下，我们可以得到面板数据——在不同时间点对于同一个观测对象展开的一系列横截面的调查——因此可以使用联合回归模型。Stimson(1985)介绍了联合回归分析；Hsiao(1986)则提供了更高级的解释。

[19] 更多关于高度多重共线性影响的详细讨论，参见 Gujarati(1988:288—298)、Johnson(1987:265—268)、Berry & Feldman(1985:40—42)以及 Hanushek & Jackson(1977:86—91)。

[20] 在方程 5.3 中，给定关于 $\varepsilon_j$ 的表达方式，假设 A5 要求 $\delta_0+(\sum_{i=1}^{m}\delta_i Z_{ij})+R_j$ 与任何一个自变量都不相关。但是 $\delta_0$ 作为一个常数以及 $R_j$ 作为一个自变量与每个变量都不相关。因此，A5 实际上是要求 $\sum_{i=1}^{m}\delta_i Z_{ij}$ 与每个自变量都不相关。

[21] 当回归模型中的因变量影响一个或多个自变量时，利用多方程模型方法——例如方程 5.4 和方程 5.5 组成的方程组——更为合适。要对这种模型作出恰当的估计，必须清楚地界定这种模型。但是即使是 OLS 回归都会出现有偏和不一致的系数估计量。其他方法，例如二阶最小二乘法(2SLS)，能提供一致的估计量。关于多方程模型的界定以及界定其估计量的讨论，参见 Gujarati(1988)、Berry(1984)以及 Hanushek & Jackson(1977)。

[22] 对于把变量包括在模型内而导致的设定误差(即，当一个变量并不在参考框架模型中，但却被包含在估计模型中时)并不需要过多关注，因为很容易避免出现这种设定误差。接下来的例子就是证明：如果一个在总体中对因变量没有影响的自变量被包含在估计模型中，那么这一变量的估计系数的期望值就为 0，同时基于样本的估计量很有可能让研究人员无法拒绝系数为 0 的原假设(当然，存在一种特殊的样本，即使错误地将变量纳入估计模型，仍会导致偏斜率系数估计量显著地不等于 0)。然而，将一个不相关的变量纳入估计模型中，还是会导致 OLS 模型中那些有关的自变量也得到无效的估计量(Berry & Feldman, 1985:18—20; Deegan, 1976; Gujarati, 1988:404—405; Maddala, 1992:164—165)。

[23] 这些参数估计量的期望值可以通过运行方程 3.1 中的 OLS 回归模型，并利用 134 名女性的总体数据得到，但是需要将 METABOLISM 这一变量排除。

[24] 对于其他实质的样本，即研究人员可以根据被排除的设定误差得到关于偏误的方向的合理推断，参见 Griliches(1957)、Kmenta(1986:446)以及 Maddala(1992:163—164)。

[25] 关于解决多重共线性问题的讨论，参见 Kelejian & Oates(1989:209—210)、Gujarati(1988:302—303)以及 Berry & Feldman(1985:47—48)。

[26] 关于这种多重共线性问题的进一步分析以及其他关于嵌套模型的方法，参见 Gujarati(1988: 413—415)。

[27] 与前文中对被排除在外的设定误差分析一致的是，当所有被排除在外的变量与每个被包括在内的变量都不相关时，偏斜率系数的估计量都是无偏的。

[28] 这一逻辑类似于多元回归中的情况。

[29] Greene(1990:772)、Dubin 与 Rivers(1989—1990:364—366，387—388)指出，这一结论非常"普遍"，但是他们也讨论了满足这种条件下的集中情况。Dubin 和 Rivers 实际上得出了结论。

[30] 关于这些名词的解释，参见注[6]。

[31] 对于一位政治学学者提出的论战，他的研究中的很多概念都被其他学者研究过(例如，投票者的偏好、政党的身份)。这些概念实际上都不是连续的，因此，回归分析不总是合适，参见 King (1989)。

[32] 我们可以构造一种方法来测量 1 英尺中最小的长度，并以此来说明由于分类导致的测量误差——这一问题我们将在本章的后半部分展开讨论。

[33] 定性的变量也被称为"定类变量"或者"分类变量"。

[34] 文献中的术语会有微妙的差异。很多人用定性变量指有两个或者两个以上类别的无序的离散变量，因此二分变量就成了其中一种定性变量。同时，另一些人用"定性的"来表示有序或者无序的离散变量。

[35] 所有的回归量都是二分变量的模型被称为"方差分析"。

[36] 因此，因变量可能为 0 或者为 1 的回归模型被称为"线性概率模型"(Aldrich & Nelson，1984: 12—19)。

[37] 普罗比和 logit 适用于当因变量为离散的模型，且有 3 个或者更多取值时；当因变量的取值没有顺序时，可以使用 logit 模型，而普罗比模型更适用于因变量为有序变量的模型。

[38] 更清楚地说，当满足 $\delta = 0$、$\mu = 1$、$E(v_j \mid T_j) = 0$ 以及 $COV(T_j, v_j) = 0$ 时，一种典型的随机测量误差模型可以被写成 $I_j = \delta + \mu T_j + v_j$ (参见 Carmines & Zeller，1979:30—32；Gujarati，1988:416；Namboodiri，Carter & Blalock，1975:539)。

[39] 这是因为当 $Y'_j = \delta + \mu T_j + v_j$(当注[38]的假设满足时)、$COV(\varepsilon_j, v_j) = 0$ 以及 $Y'$ 被用来测量方程 2.2 中的因变量时，原方程可得：

$$\mu Y_j = (\alpha - \delta) + \beta_1 X_{1j} + \beta_2 X_{2j} + \cdots + \beta_k X_{kj} + (\varepsilon_j - v_j)$$

其中误差项为 $\varepsilon_j - v_j$。如果方程 2.2 满足高斯-马尔科夫假设，本方程也满足。实际上，$\varepsilon$ 和 $v$ 结合成为一个干扰项，就相当于方程 2.2 中的 $\varepsilon$ 一样。

[40] $E(b) = \beta \cdot r_{XX}$，其中 $r_{XX}$ 指的是指标 $X$ 的信度(Berry & Feldman，1985:29)。

[41] 当然，当调查数据用 NRME 创建指标时，也可能产生一些随机测量误差。例如，如果我们说所有人都把她们的真实体重报低了整整 10%，这一假设很有可能不成立。更令人信服的假设是，人们有把她们的体重报低 10%的倾向。但是调查所得的数据还受到随机干扰项 $v$ 的影响，根据方程：

$$\text{WEIGHT}'_j = (0.90)\text{WEIGHT}_j + v_j$$

其中 $E(v_j \mid \text{WEIGHT}_j) = 0$，同时 $COV(\text{WEIGHT}_j, v_j) = 0$。有目的地误报体重会导致非随机的测量误差，然而体重测量值在某种程度上的误读或者记不清体重都会导致随机误差的出现。

[42] Namboodiri 等人(1975:579—581)把这称为"由于分类导致的测量误差"。

[43] 在线性概率模型中(见注[36])，自变量的 $X_i$ 偏斜率系数可以解释为，所有其他的自变量保持恒

定，当 $X_i$ 增加一个单位时，$Y=1$ 出现的概率的变化。

[44] 第三个例子参见 Greene 的著作(1990:724)。

[45] 在检查因变量导致的测量误差时，与这个图解反映出类似的情况，Tobit 模型更加适合回归模型(参见 Amemiya，1984；McDonald & Moffitt，1980；Tobin，1958)。在 Tobit 模型的一个应用中，一个模型的连续因变量 $Y$(例如，对于该项事业的支持度)，被假设是不可知的。因此就用一个指标 $M$(例如，贡献的大小)来指代 $Y$。其中 $M_j$ 等于 $Y_j$，$Y_j$ 比常数 $c$ 大，但是当 $Y_j \leqslant c$ 时，$M_j$ 等于 $c$。Tobit 模型也可用来处理那些被认为是"高端"或者"占据两端"的变量。

[46] 绝大多数作者涉及的模型本质上是线性的，加上"可加的"这一术语会让这个短语在意义上更加精确。

[47] 一个包括一次或多次幂变量的模型被称为"多项式模型"。

[48] 为了确定方程 5.16 模型的基本形式，我们让 $r=k=1$、$f$ 和 $h$ 作为区别方程(即，为它们贴一个"标签")以及 $g_1(X_1)=1/X_1$。

[49] 对于非线性以及非可加模型(关于总体参数是线性的和可加的)的函数形式的讨论，参见 Johnson 等(1987:239—255)、Berry & Feldman(1985:57—71)以及 Hanushek & Jackson(1977:96—101)。

[50] 对于科布-道格拉斯方程的详细讨论，参见 Gujarati(1988:189—192)以及 Kmenta(1986:511—512)。

[51] 也就是说，一个常数是从所有自变量的得分中被加上(或者被减去)的，那么变量的均值就可以被转化为 0。

[52] 设 $X$ 的均值为 0，统计学家需要认识到 $E(X^3)/E(X^2)$ 等价于 $m_3/m_2$，其中 $m_2$ 和 $m_3$ 是 $X$ 在它的均值周围的第二个和第三个时刻。

[53] 一个分布偏态若被准确测量，等于这一分布的三阶矩，即以分布的均值除以分布标准差的立方。

[54] 这是因为一个对称分布的偏态等于 0。方程的分子把偏态定义为变量的均值为 0，即 $E(X^3)$。

[55] 回想我们在方程 3.2 中计算的斜率。

[56] 方程 3.2 确定了脂肪摄入量和体重期望值关系的斜率，当 FAT = 0 时，$\beta_F+(2\cdot\beta_{FF}\cdot 0)=\beta_F$。

[57] 实际上，从本章将会讨论到的异方差性来看，我们还必须考虑的是，自变量的影响会依赖于那些对因变量产生影响但被排除在回归方程之外的变量取值的概率。

[58] 以下的处理办法由 Theil(1971:161)改编。

[59] 对于任意 $k$ 个随机变量 $W_1$，$W_2$，…，$W_k$ 的集合，如果 $a_1$，$a_2$，…，$a_k$ 为常数，那么可以得到 $E(a_1W_1+a_2W_2+\cdots+W_k)=a_1E(W_1)+a_2E(W_2)+\cdots+a_kE(W_k)$。

[60] 对于任意两个均值为 0 的变量 $X_1$ 和 $X_1$，$COV(X_1, X_2)=E(X_1\cdot X_2)$。也就是说，这两个均值都为 0 的变量的协方差等于它们乘积的期望值。

[61] 对于那些在几年的研究中连续被观测好几个星期的人而言，方程 3.1 中的一些变量可能倾向于恒定。例如，SMOKER 可能是由于这个原因而被排除在外的。其他的变量可能是因为测量方法的不同(例如，每天的食物摄入量可能是通过上个星期而非去年的摄入量来衡量的)。

[62] 这是一个不可靠的假设。例如，健康应该与脂肪摄入量呈负相关关系。同时，如果 SMOKER 随着分析阶段的不同而变化，那么这一变量与健康也相关。

[63] 这一例子来源于 Wolf(1989)关于同性恋神职人员的研究。

[64] 对于自相关的检测也基于对 OLS 回归残差的分析(Gujarati，1988:368—379)。Durbin-Watson检验是最常用的。对于这种方法的讨论及其局限性，参见 Gujarati(1988:375—379)、Kennedy(1985:100—102，105—106)、Johnson(1987:311—313)以及 Hanushek & Jackson(1977:164—168)。

[65] Sigelman 和 Dometrius(1988)提供了另一种能够解决异方差性问题的策略。他们主张在二分回归中,用 Beyle 指数对正规管辖权力的测量以及用 Abney-Lauth 测量得到的地方长官的影响的异方差性,是将"非正式的政治资源"排除在外的结果。这是因为随着地方长官处理事务能力的增长,地方长官的正式管辖权力的实际影响力应该比"非正式的"政治资源更强大。

[66] 在高阶的自回归过程中,现时出现的误差项的取值至少部分是由前两个或者更早的阶段导致的(Hibbs, 1974; Ostrom, 1978:74—76)。

[67] 对于自相关形式的详细讨论,参见 Hibbs(1974)和 Ostrom(1978)。但是 King(1989:185—187)在讨论中指出,在社会科学中,自相关常常以一阶自回归的假设形式出现是不合理的。

[68] 经历了一系列主要的、负的误差项之后,在时间 $t''$ 上数值较大的正的误差项的出现,通常可能是因为一项正的、数值较大的随机变量 $u$ 在 $t''$ 时出现。

[69] 在异方差性的情况下,OLS 估计量能够通过一系列被称为"加权最小二乘法"(WLS)的方法决定,通过将原始回归方程转换为一个有同方差的误差项的方程,然后对已转换的方程进行 OLS 回归来得到原方程的系数估计量(Berry & Feldman, 1985:87—88; Gujarati, 1988:322—325, 337—338; Wonnacott & Wonnacott, 1979:195—197)。

[70] 然而,我们可以利用异方差模型为一个 OLS 估计方程的标准计算出一个合适的替代估计量(Greene, 1990:195—197)。

[71] 当自相关以这种形式展现时,即正的一阶自回归形式,OLS 估计量的标准差的常规估计量的偏误是负的,因此,计算出来的置信区间会比真实的形式更狭窄。

[72] 这种检测依赖于 Wald 检测,参见 Greene(1990:135, 329)以获得详细的信息。

# 参考文献


Aldrich, J. H., and Nelson, F. D. (1984) *Linear Probability, Logit, and Proit Models*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—045. Beverly Hills, CA: Sage.

Amemiya, T. (1984) "Tobit models: A survey." *Journal of Econometrics* 24:3—61.

Berry, W. D. (1984) *Nonrecursive Causal Models*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—037. Beverly Hills, CA: Sage.

Berry, W. D., and Feldman, S. (1985) *Multiple Regression in Practice*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—050. Beverly Hills, CA: Sage.

Carmines, E. G., and Zeller, R. A. (1979) *Reliability and Validity Assessment*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—017. Beverly Hills, CA: Sage.

Carter, L. (1971) "Inadvertent sociological theory." *Social Forces* 50:12—25.

Deegan, J., Jr. (1976) "The consequences of model misspecification in regression analysis." *Multivariate Behavioral Research* 18:360—390.

Dye, T. R. (1966) *Politics, Economics, and the Public: Policy Outcomes in the American States*. Chicago: Rand McNally.

Fox, J. (1984) *Linear Statistical Models and Related Methods*. New York: John Wiley.

Garrow, J. S. (1974) *Energy Balance and Obesity in Man*. New York: North-Holland.

Greene, W. H. (1990) *Econometric Analysis*. New York: Macmillan.

Griliches, Z. (1957) "Specification bias in estimates of production functions." *Journal of Farm Economics* 39:8—20.

Gujarati, D. N. (1988) *Basic Econometrics* (2nd ed.). New York: McGraw-Hill.

Hanushek, E. A., and Jackson, J. E. (1977) *Statistical Methods for Social Scientists*. New York: Academic Press.

Hibbs, D. A., Jr. (1974) "Problems of statistical estimation and causal inference in time-series regression models," in H. Costner(ed.) *Sociological Methodology*, 1973—1974. San Francisco: Jossey-Bass.

Hoel, P. G. (1962) *Introduction to Mathematical Statistics* (3rd ed.). New York: John Wiley.

Hsiao, C. (1986) *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Jaccard, J., Turrisi, R., and Wan, C. K. (1990) *Interaction Effects in Multiple Regression*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—072. Beverly Hills, CA: Sage.

Johnson, A. C., Jr., Johnson, M. B., and Buse, R. C. (1987) *Econometrics: Basic and Applied*. New York: Macmillan.

Judge, G. G., et al. (1985) *Theory and Practice of Econometrics* (2nd, ed.). New York: John Wiley.

Kelejian, H. H., and Oates, W. E. (1989) *Introduction to Econometrics* (3rd ed.). New York:

Harper & Row.

Kennedy, P. (1985) *A Guide to Econometrics* (2nd ed.). Cambridge: MIT Press.

King, G. (1989) *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Cambridge: Cambridge University Press.

Kmenta, J. (1986) *Elements of Econometrics* (2nd ed.). New York: Macmillan.

Lewis-Beck, M. S. (1980) *Applied Regression: An Introduction*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—022. Beverly Hills, CA: Sage.

Luskin, R. C. (1991) "Abusis non tollit usum: Standardized coefficients, correlations and $R^2$ s." *American Journal of Political Science* 35:1032—1046.

Maddala, G. S. (1992) *Introduction to Econometrics* (2nd ed.). New York: Macmillan.

MoDonald, J. F., and Moffitt, R. A. (1980) "The uses of tobit analysis." *Review of Economics and Statistics* 62:318—321.

Mohr, L. B. (1990) *Understanding Significance Testing*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—073. Newbury Park, CA: Sage.

Namboodiri, N. K., Carter, L. F., and Blalock, H. M., Jr. (1975) *Applied Multivariate Analysis and Experimental Designs*. New York: McGraw-Hill.

Odland, J. (1988) *Spatial Autocorrelation*. Newbury Park, CA: Sage.

Ostrom, C. W., Jr. (1978) *Time Series Analysis: Regression Techniques*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—009. Beverly Hills, CA: Sage.

Pryor, F. L. (1968) *Public Expenditures in Communist and Capitalist Nations*. Homewood, IL: Irwin.

Rao, P., and Miller, R. L. (1971) *Applied Econometrics*. Belmont, CA: Wadsworth.

Romieu, I., et al. (1988) "Energy intake and other determinants of relative weight." *American Journal of Clinical Nutrition* 47:406—412.

Schroeder, L. D., Sjoquist, P. L., and Stephan, P. E. (1986) *Understanding Regression Analysis: An Introductory Guide*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—057. Beverly Hills, CA: Sage.

Sigelman, L., and Dometrius, N. C. (1988) "Governors as chief administrators: The linkage between formal powers and informal influence." *American Politics Quarterly* 16:157—170.

Stimson, J. A. (1985) "Regression in space and time: A statistical essay." *American Journal of Political Science* 29:914—947.

Sudman, S. (1976) *Applied Sampling*. New York: Academic Press.

Theil, H. (1971) *Principles of Econometrics*. New York: John Wiley.

Tobin, J. (1958) "Estimation of relationships for limited dependent variables." *Econometrica* 26:24—36.

Wolf, J. (1989) *Gay Priests*. San Francisco: Harper & Row.

Wonnacott, R. J., and Wonnacott, T. H. (1979) *Econometrics* (2nd ed.). New York: John Wiley.

## 译名对照表

| | |
|---|---|
| additivity | 可叠加性 |
| analysis of variance | 方差分析 |
| autocorrelation | 自相关 |
| auxiliary regression | 辅助回归 |
| Best Linear Unbiased Estimators(BLUE) | 最优线性无偏估计量 |
| Cobb-Douglas function | 科布-道格拉斯方程 |
| disturbance term | 干扰项 |
| efficiency | 有效性 |
| error term | 误差项 |
| expected value | 期望值 |
| Gauss-Markov assumptions | 高斯-马尔科夫假设 |
| generalized least squares(GLS) | 广义最小二乘法 |
| heteroscedasticity | 异方差性 |
| homoscedasticity | 同方差性 |
| intercept errors | 截距误差 |
| linear probability model | 线性概率模型 |
| linearity | 线性 |
| maximum likelihood method | 最大似然估计法 |
| measurement error | 测量误差 |
| multicollinearity | 多重共线性 |
| nonlinearity | 非线性 |
| Nonrandom Measurement Error(NRME) | 非随机测量误差 |
| normal distribution | 正态分布 |
| ordinary least squares(OLS) | 普通最小二乘法 |
| parameters | 参数 |
| partial slope coefficients | 偏斜系数 |
| probit model | 普罗比模型 |
| proxy variables | 代理变量 |
| random measurement error(RME) | 随机测量误差 |
| residuals | 残差 |
| scale error | 尺度误差 |
| selection bias | 选择性偏误 |
| serial correlation | 序列相关 |
| skewness | 偏态 |
| specification error | 设定残差 |

| | |
|---|---|
| time-series regression model | 时间序列回归模型 |
| truncated sample | 截断样本 |
| two-stage least squares(2SLS) | 二阶最小二乘法 |
| unbiasedness | 无偏性 |
| weighted least squares(WLS) | 加权最小二乘法 |

# 回归诊断简介

## 作 者 简 介

### 约翰·福克斯(John Fox)

加拿大多伦多市约克大学社会学、数学和统计学教授，并担任社会调查研究所统计咨询服务协调人。目前的研究工作主要包括统计学方法与政治经济学研究，他也是《线性统计模型和相关方法》(Wiley, 1984)的作者与《现代数据分析方法》(Sage, 1990)的联合主编。

## 译 者 简 介

### 於嘉

2009年获中国人民大学法学学士学位，现为香港科技大学社会科学部硕士研究生。

# 序

在社会科学的数据分析中，回归可谓最常用的方法。通过计算机获得一个估计的回归方程就和数 1、2、3 一样简单，事实的确如此，因为利用任何一个软件程序，研究者都可以按如下步骤操作：(1)加载样本数据；(2)确定回归方程；(3)利用普通最小二乘法进行估计。这将获得一个类似下面这下等式的结果：

$$Y = 62 + 71.5X_1 + 5.4X_2 + e$$

但是，这个估计的结果如实反应了真实世界的状况吗？例如，在 $X_2$ 保持不变的情况下，$X_1$ 一个单位的变化是否将导致 $Y$ 产生 71.5 的预期变化？我们往往可以非常自信地谈论总体估计的精确度。但是，我们对回归结果的信任程度取决于是否能够成功地处理以下常见问题：多元共线性、奇异值、非正态、异方差性以及非线性。

Fox 教授将“诊断”引申为发现上述问题。例如奇异观测值或更概括地讲，即强影响观测值产生的问题。除了那些可以展示某一极端值如何影响回归直线的常用图形外，他对其他测量方法也进行了阐释：预测值、学生残差、Cook 距离以及偏回归散点图。这些测量方法大多可以通过常用的软件程序获得，例如 SAS 或 SPSS。

在对回归进行了诊断之后，Fox 专注寻找可能的解决办法。此类问题非常多，例如，如果具有高度的共线性，这个变量需要被剔除出回归方程吗？如果有奇异值出现，这个观测是否应该被舍弃？当误差的分布是偏斜的时候，是否应该对其进行一些变换？在异方差性存在的情况下，是否应该使用加权最小二乘法以解决这一问题？当非线性问题存在时，是否应该使用次方转换？在面对这些重要的问题时，应尽量避免使用机械的权宜方法。正如作者不断强调的，这些方法永远不能取代判别和思想。

为了使解释更加丰富，Fox 利用了许多数据作为例子：美国的人口普查、职业声望、人们报告的体重、加拿大公司中的董事会。这些例子使得本书中的诊断适用于广大的回归方法使用者。此外，有意愿受更高级训练的读者可以在附录中寻找答案(例如，对用于解决高度共线性的岭回归的评估)。每一个使用回归分析的人，理应进行一系列回归诊断。

迈克尔 · S. 刘易斯-贝克

# 第 1 章 | 概论

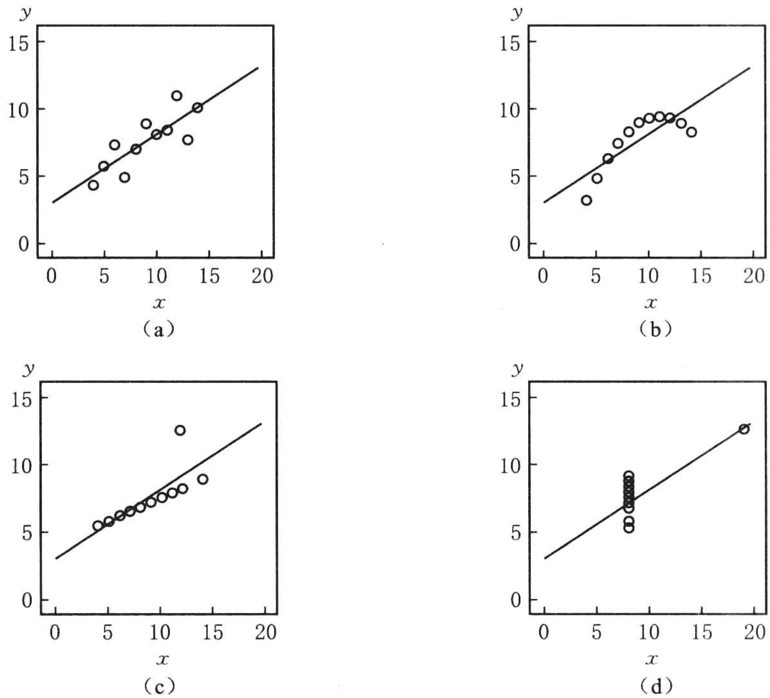
在社会科学研究中，线性最小二乘回归分析可谓最常用的统计技术，并为许多其他的统计方法奠定了基础。但是，最小二乘回归往往面临许多困难，它对于数据结构有着较强且往往不切实际的假设。回归诊断是用于探索存在于回归分析中的问题及判断某些假设是否合理的一种技术。

回归诊断在当代的发展与计算机交互式的统计分析的实现是不可分割的，因此，回归诊断在很大程度上是近 20 年的产物。与回归诊断方法紧密相关的是用于纠正已发现问题的各种技术，其中许多方法都涉及对数据的转换。

作为一个初步的例子，我们首先考虑图 1.1 中来自 Anscombe(1973)的 4 幅散点图。统计分析的一个目的就在于为数据提供详尽的描述性归纳。Anscombe 的 4 个数据集已被设计得出相同的标准线性回归结果：斜率、截距、相关系数、回归标准误、系数标准误以及统计检验。但非常重要的是，它们不具有相同的残差。

在图 1.1(a)中，线性回归合理地描述了 $y$ 随 $x$ 的增长而增长这一趋势。在图 1.1(b)中，线性回归未能反映出数据具有的曲线形式，所以线性方程显然是错的。在图 1.1(c)中，某一点与其他点构成的直线偏离，这对拟合的回归直线产生了很大的影响，而仅通过其他点的直线则将完美拟合。在理想的情况下，我们希望了解为什么最后一个观测值偏离了其他观测值。它可能确有特殊之处(例如它受到除 $x$ 之外其他值的影响)，或是体现了在数据记录过程中的误差。当然，我们在此只是设想，因为 Anscombe 的数据只是简单地构造出来的，但重点在于我们需要从实质上寻求解决异常值的方法。在图 1.1(d)中，若没有最后一个点，我们就不能拟合出直线。因此，我们至少应该对回归结果持有谨

慎而怀疑的态度。




注:来自 F. J. Anscombe(1973)。在每个散点图中都显示了最小二乘回归直线。

资料来源:Anscombe(1973)。已经获得美国统计协会授予的重新作图和印刷权。

**图 1.1　具有同样标准回归输出的数据集**

上述例子的简单阐释一定程度上介绍了本书的许多议题,包括非线性、奇异数据、强影响数据以及图示的重要作用。普通的回归结果往往不能清晰地表现出全部的真实状况,诊断方法(其中许多是使用图形的)帮助我们填补回归结果未顾及的那些部分。

第 2 章主要回顾最小二乘线性回归。第 3 章将讨论多元回归中共线性的问题。第 4 章主要处理奇异与强影响数据。在第 5 章至第 7 章中,我们将主要探讨误差非线性、不一致的误差方差和非线性问题。第 8 章将简要阐释离散数据产生的问题。第 9 章主要介绍基于最大似然法、计分检验和构造变量的较复杂的诊断方法。在第 10 章中,我们将考虑如何将介绍的具体诊断方法和技术应用

到研究中。这一章的内容也包括如何使用电脑软件进行回归诊断，并以阅读书目的推荐作为全书的结束。

大部分技术性的细节可参见本书的附录。对附录内容理解的基本要求是掌握一定的最小二乘法的矩阵代数以及基本的统计理论。尽管附录提供了更加深入的内容，但不十分重视技术的读者仍可以略过，这并不影响对本书的理解。我的目标在于使这本书在一定程度上是独立的，同时保持其通用性。

本书当然无法包含有关诊断的全部内容，但我试图处理一些可以使回归模型更加有效的中心议题。由于篇幅限制，除了在第 10 章中略有提及，本书不包括解决时间序列回归中的误差自相关问题的方法。关于这方面的内容，可参见另一本专著(Ostrom，1990)。

# 第 2 章 | 最小二乘回归

由于我们已经假定本书的读者对于最小二乘回归非常熟悉，因此本章的主要目的在于对其进行回顾。在附录 1 中有具体演算过程。

## 回归模型

线性回归模型可以用以下方程表示出来：

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad [2.1]$$

其中，$i = 1, \cdots, n$ 表示样本观测。在方程 2.1 中，$y_i$ 是因变量，$x_{ij}$ 是回归因子，$\varepsilon_i$ 是不可观测的误差。$\beta_j$ 是需要从数据中估计的未知参数。按照通常的标准，我们假设误差是独立的，且符合期望为 0、方差为常数 $\sigma^2$ 的标准正态分布：$\varepsilon_i \sim \text{NID}(0, \sigma^2)$。违背这一假设的结果和用于发现是否违背这一假设的方法将在后面进行讨论。

如果 $x_{ij}$ 与 $y_i$ 都是由抽样获得的，而不像实验设计那样是限定的，则还需要假设 $x$ 的分布独立于 $\sigma$。最后一个假设既可以被认为是描述性的也可以被认为是结构性的：从描述方面来讲，任何 $x$ 值上 $y$ 所有取值的均值和 $x$ 值本身构成的点必须在回归平面上；从结构方面或因果方面来讲，我们另外要求 $y$ 被忽略的原因（其为包含在误差中的一部分）本身不受 $x$ 影响，且与 $x$ 线性不相关。除非特殊情况，否则最后一个假设是无法用数据检验的，因为最小二乘拟合确保了用于估计误差的残差与样本中的 $x$ 是不相关的。

## 最小二乘估计

拟合的回归可以写为：

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i = \hat{y}_i + e_i$$

其中，$x_{ij}$ 和 $y_i$ 与方程 2.1 中的一样，$b_j$ 是相应的 $\beta_j$ 的估计，$e_i$ 是残差。拟合的值可以通过 $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$ 获得。用于确保残差平方和最小而获得的最小二乘回归系数是符合下列标准方程的 $b_j$ 值：

$$b_0 n + b_1 \sum x_1 + \cdots + b_k \sum x_k = \sum y$$

$$b_0 \sum x_1 + b_1 \sum x_1^2 + \cdots + b_k \sum x_1 x_k = \sum x_1 y$$

$$\vdots$$

$$b_0 \sum x_k + b_1 \sum x_1 x_k + \cdots + b_k \sum x_k^2 = \sum x_k y$$

由于总数显然超过了 $i = 1, \cdots, n$，所以我停止使用用于表示观测值的下标 $i$（例如 $x_1$ 代表 $x_{1i}$）。上述标准方程对 $b_j$ 有唯一解值，但需要满足两个条件：(1)所有的 $x_j$ 都不是恒定的；(2)任何 $x_j$ 都不能是其他的完全线性组合。

标准方程显示最小二乘残差和为 0，因此其平均值也为 0。此外，这些残差和拟合值与 $x$ 均不相关，原因在于：

$$\sum e_i \hat{y}_i = 0$$

$$\sum e_i x_{ji} = 0 (j = 1, \cdots, k)$$

误差方差是根据 $s^2 = \sum e_i^2 / (n - k - 1)$ 估计得出的，其中 $n - k - 1$ 是误差的自由度。拟合模型复相关系数的平方可以表示如下：

$$R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum e_i^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

它可以解释为用 $x$ 进行线性回归来解释的 $y$ 的比例。

## 回归系数的统计推论

被估计的回归系数 $b_1$，…，$b_k$ 的抽样方差为：

$$\hat{V}(b_j)=\frac{s^2}{\sum(x_{ji}-\bar{x}_j)^2}\times\frac{1}{1-R_j^2}=\frac{s^2}{(n-1)s_j^2}\times\frac{1}{1-R_j^2}$$

其中，$s_j^2=\sum(x_{ji}-\bar{x}_j)^2/(n-1)$ 是 $x_j$ 的方差，$R_j^2$ 是通过利用其他 $x$ 对 $x_j$ 进行回归而得到的复相关系数平方。零假设 $H_0$：$\beta_j=\beta_j^{(0)}$（通常 $H_0$：$\beta_j=0$）的 $t$ 统计量是由 $t_0=(b_j-\beta_j^{(0)})/\mathrm{SE}(b_j)$ 获得的，其中 $\mathrm{SE}(b_j)=[\hat{V}(b_j)]^{1/2}$ 是 $b_j$ 的估计标准误。根据 $H_0$，$t_0$ 符合 $t$ 分布，其自由度为 $n-k-1$。

为了检验所有回归系数均为 0 这一假设（除了常数 $\beta_0$），例如，$H_0$：$\beta_1=\beta_2=\cdots=\beta_p=0$（其中 $p\leqslant k$），我们可以计算增量 $F$ 统计值：

$$F_0=\frac{n-k-1}{p}\times\frac{R^2-R_0^2}{1-R^2}$$

在这里，$R^2$ 与之前一样，是全模型复相关系数的平方值，而 $R_0^2$ 是利用剩余 $x$，即 $x_{p+1}$，…，$x_k$ 对 $y$ 进行回归获得的复相关系数的平方值。这些 $t$ 统计和 $F$ 统计均符合回归模型的假设，包括正态分布的误差。

$\beta_j$ 的 $100(1-\alpha)\%$置信区间为：

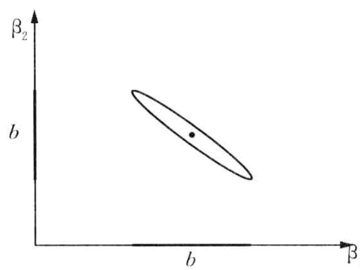
$$\beta_j=b_j\pm t_{\alpha/2,\,n-k-1}\mathrm{SE}(b_j) \qquad [2.2]$$

由于置信区间的范围与估计的系数标准误是成比例的，$\mathrm{SE}(b_j)$自然成为对估计量 $b_j$ 估计精度的一个测量。

同样，一个椭圆形的多系数联合置信区域可以通过回归系数的方差和协方差以及来自 $F$ 分布的一个临界值获得（见附录 1）。图 2.1 显示了参数个数为两个（$\beta_1$ 与 $\beta_2$）的状况。正如方程 2.2 给出了在 $\alpha$ 水平下所有可接受的 $\beta_j$ 值，图 2.1 中的椭圆形包括了所有 $\beta_1$ 与 $\beta_2$ 联合的可接受值。

置信椭圆形以估计值 $b_1$ 与 $b_2$ 为中心。椭圆在 $\beta_1$ 与 $\beta_2$ 轴上的投影表示每

个参数单独的置信区间，尽管与联合区域相比，投影往往在具有高水平的置信度区间。正如置信区间的长度表示单一参数估计的精度，联合置信区域的大小（比如 2 个参数时的面积，3 个参数时的体积和 4 个参数时的多维体积）表示这几个参数联立的估计精度。




注：置信区域组成的椭圆形是以估计量 $b_1$ 与 $b_2$ 为中心的。联合置信区域在 $\beta_1$ 和 $\beta_2$ 轴上的投影即这些参数的置信区间（但是与联合置信区域相比，置信水平较高）。

**图 2.1　系数 $\beta_1$ 和 $\beta_2$ 的联合置信区域**

## 一般线性模型

由于除了不能与误差相关之外，没有关于 $x$ 分布的假设，这使得线性回归模型的应用范围远比其最初出现时广泛得多。$x$ 可能包括构造的、用以获得定性自变量作用的虚拟回归因子；或由几个变量构成的、用以了解自变量的非累加作用的交互回归因子；或用于对数据中的非线性形式建模的多项式回归因子等等。只要模型可以表示为方程 2.1 所示，也就是说，模型对参数 $\beta_0$，$\beta_1$，…，$\beta_k$ 是线性的，那么就可以用线性回归进行分析，回归平面自身不必是平坦的。广义来说，线性回归模型也可称为“一般线性模型”。

# 第 3 章 | 共线性

## 共线性与方差膨胀

正如在第 2 章中提到的，当线性回归模型的回归因子中存在完全的线性关系时，最小二乘系数将不是唯一确定的。这一结果可以从只具有两个回归因子，即 $k = 2$ 的情况下看出，其标准方程如下：

$$\begin{aligned} b_0 n + b_1 \sum x_1 + b_2 \sum x_2 &= \sum y \\ b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 &= \sum x_1 y \\ b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 &= \sum x_2 y \end{aligned} \tag{3.1}$$

解这个标准方程可以获得：

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 \\ b_1 &= \frac{\sum x'_1 y' \sum x'^2_2 - \sum x'_2 y' \sum x'_1 x'_2}{\sum x'^2_1 \sum x'^2_2 - (\sum x'_1 x'_2)^2} \\ b_2 &= \frac{\sum x'_2 y' \sum x'^2_1 - \sum x'_1 y' \sum x'_1 x'_2}{\sum x'^2_1 \sum x'^2_2 - (\sum x'_1 x'_2)^2} \end{aligned} \tag{3.2}$$

其中，$x'_1 = x_1 - \bar{x}_1$，$x'_2 = x_2 - \bar{x}_2$，$y' = y - \bar{y}$ 是均差形式的变量。

$x_1$ 与 $x_2$ 的相关系数为：

$$r_{12} = \frac{\sum x'_1 x'_2}{\sqrt{\sum x'^2_1 \sum x'^2_2}}$$

因此，如果 $r_{12}=\pm 1$，则方程 3.2 中 $b_1$ 与 $b_2$ 的分母为 0，这些系数无解（更准确地说，$b_1$ 与 $b_2$ 有无限组符合标准方程 3.1 的解）。

如果自变量之间具有很强的但不完全的线性关系，则会导致最小二乘回归系数的不稳定：系数的标准误很大，反映了 β 值估计的不准确，因此，β 的置信区间范围也很大。数据中的变化，哪怕是极端情况下因为四舍五入导致的误差，也可以在很大程度上改变最小二乘系数，而且由于最小二乘值而导致的系数的较大变化也很难增加残差的平方和。

在前面的章节中，我们提到最小二乘系数 $b_j$ 的估计方差为：

$$\hat{V}(b_j)=\frac{s^2}{(n-1)s_j^2}\times\frac{1}{1-R_j^2} \qquad [3.3]$$

$1/(1-R_j^2)$ 表示了共线性对估计精度产生的影响大小，叫做"方差膨胀因子"（$\text{VIF}_j$）。需要注意的是，VIF 显示的不是两个回归因子之间的相关性（当 $k>2$ 时），而是对某一个自变量 $x$ 根据其他所有自变量进行回归得到的复相关系数。因此，多元回归中的共线性也被称为"多元共线性"。

另外值得关注的是，在方程 3.3 中，影响估计精度的其他因素是估计的误差方差、样本规模和 $x_j$ 的方差。误差方差越小，样本规模越大；$x$ 的分布越广，则回归的估计精度越高。从已有的经验来看，社会科学研究中不精确的估计大多来自过大的误差方差和过小的样本规模，而不是严重的共线性。

由于 $\beta_j$ 的估计精度可以用参数置信区间的宽度来衡量，并且由于置信区间的宽度与 $\beta_j$ 的标准误是成比例的，我推荐对 VIF 的平方根而非 VIF 本身进行检查。表 3.1 显示，$x$ 间的线性关系必须非常强，才能对回归有严重的影响。例如，只有当 $R_j$ 接近 0.9 的时候，估计的精度才会减半。

以 Ericksen、Kadane 和 Tukey 的数据为例，其回归结果显示在表 3.2 中。此处的目标在于创建一个预测方程以提高对 1980 年美国人口普查不完全统计的估计。我们已经能够肯定，人口普查未能对每个郡中的所有人进行调查。

**表 3.1　回归因子间的复相关函数构成的系数方差膨胀**

| $R_j$ | $\mathrm{VIF}_j = 1/(1-R_j^2)$ | $\sqrt{\mathrm{VIF}}$[a] |
|---|---|---|
| 0.0 | 1.0 | 1.0 |
| 0.2 | 1.04 | 1.02 |
| 0.4 | 1.19 | 1.09 |
| 0.6 | 1.56 | 1.25 |
| 0.8 | 2.78 | 1.67 |
| 0.9 | 5.26 | 2.29 |
| 0.95 | 10.3 | 3.20 |
| 0.99 | 50.3 | 7.09 |
| 0.999 | 500.3 | 22.4 |
| 1.0 | ∞ | ∞ |

注：a. 标准误差 $b_j$ 的影响。

**表 3.2　对美国 66 个中心城市、州城市和州 1980 年进行的人口普查不完全统计的估计进行的回归**

| 预测因子 | 系　数 | 标准误 | $\sqrt{\mathrm{VIF}}$ |
|---|---|---|---|
| 常数项 | −1.77 | 1.38 | — |
| 少数族群 | 0.0798 | 0.0226 | 2.24 |
| 犯罪 | 0.0301 | 0.0130 | 1.83 |
| 贫困 | −0.178 | 0.0849 | 2.11 |
| 语言 | 0.215 | 0.0922 | 1.28 |
| 高中 | 0.0613 | 0.0448 | 2.15 |
| 住房 | −0.0350 | 0.0246 | 1.37 |
| 城市 | 1.16 | 0.77 | 1.88 |
| 便利 | 0.0370 | 0.0093 | 1.30 |
| $R^2$ | 0.708 | | |

注：这些作者使用了权重最小二乘回归（参见附录 8），由此可以考虑最初对 66 个地区不完全统计估计的不同精度。与之相比，上表显示的是普通最小二乘回归。

某些类型的个体更容易被漏查，比如非白种人、穷人和大城市中的居民。回归中的因变量是对 66 个地区进行人口普查不完全统计状况的初步估计。这 66 个地区包括 16 个大城市、这 16 个大城市所在州的剩余地区以及另外 34 个州。初步的估计是根据 8 个被认为会对人口普查不完全统计构成影响的预测因子进行回归得出的，这 8 个自变量如下：

(1) 黑人或西班牙裔所占的百分比(少数族群);

(2) 每 1000 人中发生严重犯罪的比例(犯罪);

(3) 贫困群体的百分比(贫困);

(4) 英语口语与写作有障碍群体人数百分比(语言);

(5) 25 岁及以上没有高中毕业者所占人口百分比(高中);

(6) 住房较小或多单元住宅群体所占人口百分比(住房);

(7) 城市为 1、州及州内其他城市为 0 的虚拟变量(城市);

(8) 能够便利地接受访问与拒绝寄回问卷户数的百分比(便利)。

表 3.3 显示了这些预测因子之间的相关系数。尽管某些成对的相关系数相对较大(最大的接近 0.73),但并没接近 1 的。从表 3.2 中可以看到,根据 VIF 的平方根,几个回归估计(少数族群、贫困和高中的系数)受共线性的影响。

**表 3.3 对 1980 年人口普查不完全统计的 8 个预测因子之间的相关系数**

| 预测因子 | 少数族群 | 犯罪 | 贫困 | 语言 | 高中 | 住房 | 城市 |
|---|---|---|---|---|---|---|---|
| 犯罪 | 0.655 | | | | | | |
| 贫困 | 0.738 | 0.369 | | | | | |
| 语言 | 0.395 | 0.512 | 0.152 | | | | |
| 高中 | 0.535 | 0.0666 | 0.751 | −0.116 | | | |
| 住房 | 0.356 | 0.532 | 0.335 | 0.340 | 0.235 | | |
| 城市 | 0.758 | 0.729 | 0.538 | 0.480 | 0.315 | 0.566 | |
| 便利 | −0.334 | −0.233 | −0.157 | −0.108 | −0.414 | −0.0863 | −0.269 |

资料来源:Ericksen, Kadane & Tukey(1980)。

作为直接衡量共线性对估计精度影响的指标,系数方差的膨胀度可以扩展用于几个系数的置信区域。相关的应用包括虚拟变量或多项式变量的情况,但在这里,单一系数的方差膨胀则不太受关注。

## 对共线性的处理:没有速效方法

当 $x_1$ 与 $x_2$ 之间共线性问题很突出时,例如在 $x_2$ 统计被控制恒定的情况下,

数据不包括任何由 $x_1$ 带来的影响，因为当 $x_2$ 被固定时，$x_1$ 也没有任何变化。当然，当固定 $x_1$ 时，$x_2$ 的情况也一样。因为 $b_1$ 估计了 $x_2$ 固定时，$x_1$ 的局部效应。尽管有许多用来处理共线性的方法，但没有一个能从数据中提取出根本不存在的信息，否则就是研究的问题被不经意间重新定义了。在一些情况下，这种重新定义是有理可循的，但通常情况下并非如此。解决共线性问题最理想的方式就是在避免类似问题的情况下收集新的数据，例如对 $x$ 进行实验操作。但不幸的是，这个解决方法往往不切实际。

有几种不能充分解决共线性数据的处理方法将在下文进行讨论。此处我用较大的篇幅探讨变量选择问题，因为选择的方法往往已被社会科学家滥用，并且关于变量选择的策略比较直接，另外，在某种或有限制的情况下，变量选择往往是一种合理的解决对策。

第一，模型的重新确定。尽管共线性是数据中的问题而非（必然）模型的缺陷，但一种解决此类问题的方法是模型的重新确定。也许经过进一步的思考会发现，某些模型中的回归因子可能是同一个潜在建构的、不同的概念化指标。所以这些测量可以使用某些方法进行合并，或者可以选择用来表示其他的建构。在这种情形下，被研究的自变量 $x$ 的高度相关恰恰显示了高度的有效性。假设一个跨国分析是针对影响婴儿死亡率的因素，那么自变量中的人均国民生产总值、人均耗能量以及人均电视拥有率会高度相关。此时，研究可能将这些变量处理为展现总体经济发展水平的一个指标。

另外，我们可以重新考虑是否需要在检查 $y$ 与 $x_1$ 的关系时控制 $x_2$。一般来说，这一类的重新确认只发生在初始模型不理想或研究者想要放弃一些研究目标之时。例如，假设一个时间序列回归旨在检验决定已婚妇女参与劳动的决定因素，共线性问题使我们很难分离男性工资水平与女性工资水平的影响。但是在这个研究中，我们仍希望在控制其他自变量时，理解妇女的工资与参与劳动力市场的局部关系。

第二，变量选择。一个常见的但往往容易被误用的解决共线性的方法是变量选择，它往往有一定的步骤，用来将模型中回归因子减少至较低相关性的组合。向前逐步回归的方法是每次在模型中加入一个变量。在每一步中，使 $R^2$

增量最大的变量将被选择留下。这一步骤在增量比预先设定的标准小时停止。向后逐步回归方法与之类似，差别在于全部过程从全模型开始，且每次删掉一个变量。向前/向后的方法是上述两种方式的组合。

逐步的方法往往被一些不成熟的研究者滥用，他们试图将变量纳入回归方程的次序作为对这些变量重要性的解释。这类处理方式可能是误导性的，因为在某种情况下，两个高度相关的自变量 $x$ 可能对 $y$ 有同样的影响，但只有一个可以被纳入回归方程，因为另一个不能增加任何附加信息。对数据稍加处理或另选一个样本则可能导致相反的结果。

在技术上对逐步方法的反对是因为其可能无法显示出给定数量的最佳回归因子的组合子集（比如能够使 $R^2$ 达到最大的子集）。计算机技术的进步可以使我们在计算过程中检查所有回归因子的子集，即便回归因子的个数 $k$ 非常大。除了使选择的标准最优化之外，选择子集这一技术也对揭示其他可能的或几乎等同的模型大有裨益，这样可以帮助避免产生唯一“正确”的结果的状况。

一个常用的选择子集的方法是基于所有（或定额）的从 $\hat{y}$ 估计得出的 $E(y)$ 的均方差，即根据观察到的拟合平面中的 $x$ 估计总体回归平面：

$$\gamma_p = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{MSE}(\hat{y}_i) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\{V(\hat{y}_i) + [E(\hat{y}_i) - E(y_i)]^2\} \qquad [3.4]$$

其中，拟合值 $\hat{y}_i$ 是基于包括 $p \leqslant k+1$ 个回归因子的模型（包含常数项，它往往都包含在模型中）得出的。如果研究目标就是根据 $x$ 预测 $y$，那么使用误差作为估计$E(y)$的标准就是合理的。

需要注意的是，方程 3.4 中的$[E(\hat{y}_i)—E(y_i)]^2$ 表示对总体回归平面 $E(y_i)$得出的估计值 $\hat{y}_i$ 的偏差平方。当共线性的回归因子从模型中被删除的时候，一般来说，$V(\hat{y}_i)$会变小（取决于数据点的构成与真实的 $\beta$ 在回归因子中被删除），但是偏差则可能被引入拟合值。因为 MSE 是方差的和与偏差的平方，所以根本的问题在于，方差的降低是否造成了偏差的增加？

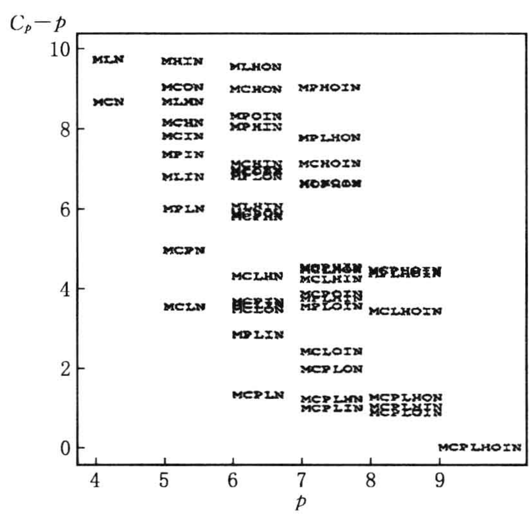
Mallow(1973)的 $C_p$ 统计量将 $\gamma_p$ 估计为：

$$C_p = \frac{\sum e_i^2}{\hat{\sigma}^2} + 2p - n = (k+1-p)(F_p - 1) + p$$

其中，残差来自考虑中的子模型，误差方差的估计量 $\hat{\sigma}^2$ 是全模型中的 $s^2$，$F_p$ 是 $F$ 统计量的增量，用以检验现在的子模型中被忽略的回归因子总体系数为 0 这一假设。如果这个假设是成立的，则 $E(C_p) \approx p$。因此，一个好的模型往往拥有接近或小于 $p$ 的 $C_p$ 值。同样，使 $C_p$ 值最小也会导致残差平方和的最小化，从而使 $R^2$ 最大化。需要注意的是，对于全模型来说，$C_{k+1}$ 必然等于 $k+1$。

由于好的模型拥有接近于 $p$ 的 $C_p$ 值，所以我们可以依据 $p$ 来对 $C_p$ 进行绘图，从而辨识出好的模型。在此图中，将每个点都用符号标示以代表包含在模型中的自变量，并将 $C_p = p$ 这条线叠加在这个散点图上，好的模型会接近或低于这条参照线。如果依据 $p$ 对 $C_p$ 进行绘图去除了趋势（即每一个点都减去参照线），那么这个散点图将非常易于观察。此时我们可以寻找 $C_p - p$ 接近或小于 0 的值。

图 3.1 是关于人口普查中不完全统计的一个解释性的 $C_p$ 散点图。图中只包含使 $C_p - p \leqslant 10$ 的模型（包括 $2^8 - 1 = 255$ 个预测因子子集中的 52 个）。Ericksen 等人（1992）选择了图中标为 MCN 的子集（包括少数族群、犯罪与便利这 3 个预测因子）。在此，$p = 4$ 且 $C_p = 12.7$。




注：大写字母用于标示变量，即少数族群（Minority）、犯罪（Crime）、贫困（Poverty）、语言（Language）、高中（High schoold）、住房（hOusing）、城市（cIty）和便利（coNventional）。Ericksen 等人（1989）选择了自变量子集 MCN（即少数族群、犯罪和便利）。

**图 3.1　对普查不完全统计的 $C_p - p$ 根据 $p$ 绘制散点图**

这表明了仍有提升的空间，有待引入更多的预测因子。表 3.4 包含了这个子集的回归方程、4 个回归因子（MCLN，加入了语言，$p=5$ 且 $C_p=8.5$）和 5 个回归因子（MCPLN，加入了贫困，$p=6$ 且 $C_p=7.3$）的"最佳"子集的回归方程。对于这个数据集来说，向后与向前/向后逐步回归的方法确认了包括 3 个、4 个和 5 个预测因子的"最佳"子集，但是向前的方法却无法做到（但是通过回顾之前的内容，Ericksen 等人采用了一种比普通最小二乘回归法更复杂的估计策略）。

**表 3.4 最优的模型回归子集**

| 预测因子 | 系数[a] | | |
|---|---|---|---|
| | $p=4$ | $p=5$ | $p=6$ |
| 常数项 | −2.22 | −1.98 | −0.793 |
| | (0.56) | (0.55) | (0.860) |
| 少数族群 | 0.0786 | 0.0752 | 0.101 |
| | (0.0147) | (0.0143) | (0.020) |
| 犯罪 | 0.0363 | 0.0272 | 0.0243 |
| | (0.0100) | (0.0104) | (0.0103) |
| 便利 | 0.0280 | 0.0273 | 0.0293 |
| | (0.0081) | (0.0077) | (0.0077) |
| 语言 | | 0.209 | 0.184 |
| | | (0.087) | (0.086) |
| 贫困 | | | −0.110 |
| | | | (0.062) |
| $R^2$ | 0.638 | 0.669 | 0.686 |
| $C_p$ | 12.7 | 8.51 | 7.32 |

注：a. 括号内为系数的标准误。
资料来源：数据来自 Ericksen，Kadane & Tukey(1989)。

在进行变量选择时，应该牢记以下的告诫：

首先，最重要的是，变量选择导致了一个重新确立的模型，但是往往不能解决我们一开始提出的研究问题。尤其在最初的模型是正确确立的，并且包含的与忽略的变量是相关的情况下，根据变量选择而获得的系数估计量将是有偏的。这样一来，这类方法最适用于纯粹的预测问题，其中根据回归因子得出的预测数据的值将在这个选择发生的数据框架之内，正如在人口普查不完全统计

的例子中。在这种情况下，即使系数本身是有偏的，仍然有可能获得 $E(y)$ 的良好估计值。但是，如果对于一个新的观测值来说，$x$ 在它与那些得出估计值的观测中是不同的，则相应的预测 $y$ 可能严重有偏。

其次，当回归因子成系列出现时（例如虚拟变量），这些系列则应在选择过程中被保留在一起。同样，当回归因子中有分层关系时，那么这些关系应被保留。比如，如果主效应不包含在回归因子中，那么包含这个主效应的交互项回归因子也不应该出现在模型中。

第三，因为变量选择使模型对样本数据的拟合最优化，所以基于变量选择的自变量系数的标准误（以及置信区间和假设检验）往往会夸大结果的精确度。因此，利用样本的偶然性特征是有风险的。关于这个问题的解决方法，我将在第 10 章对交叉效度的讨论中提及。

最后，即便是在没有严重的共线性问题时，变量选择也可用于统计建模。删除估计系数很小的回归因子往往是没有问题的，这也将建立一个更简约的模型。事实上，在一个大样本中，删除很小的但是具有“统计显著”的系数也是合理的。

另一种处理共线性数据的方法是有偏估计。这种方法的思路是用一小部分系数估计值的偏差换取系数样本方差的大幅降低。得到对 $\beta$ 的估计结果与最小二乘估计相比，拥有较小的均方误（对比关于估计回归子集的讨论）。最常见的有偏估计模型称为“岭回归”（在附录 2 中有简短的介绍）。

与变量选择一样，有偏估计对于共线性问题来说也不是神奇的万灵药。例如，岭回归涉及选择一个任意的岭常数以控制岭估计与最小二乘估计之间差异的程度：岭常数越大，偏差越大，而岭估计的方差也越小。不幸但也可以理解的是，为了选择一个最优的岭常数（哪怕只是一个好的岭常数），往往也需要一些我们试图估计的、未知的 $\beta$ 的信息。我在此提及有偏估计的意图正在于对其通常的使用提出告诫。

处理共线性数据的最后一个办法就是引入额外的预知信息，帮助降低由共线性带来的模糊状况。有几种不同的方法可以将预知信息用于回归中，但我们应该用一个简单的案例来解释这一方法。更复杂的方法将不在此处进行讨论，

这些方法有时难以应用到实践中(参见 Belsley, Kuh & Welsch, 1980:193—204; Theil, 1971:346—352)。

假设我们想估计以下模型:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

其中,$y$ 是储蓄,$x_1$ 是工资收入,$x_2$ 是个人股票收入,$x_3$ 是利息收入。假设我们很难估计 $\beta_2$ 与 $\beta_3$,因为 $x_2$ 与 $x_3$ 高度相关。进一步假设我们有理由相信 $\beta_2 = \beta_3$,并用常量 $\beta_*$ 表示。如果 $x_2$ 与 $x_3$ 不高度相关,我们可以直接对 $\beta_2 = \beta_3$ 这一点进行假设检验。在这种情况下,我们拟合模型为:

$$y = \beta_0 + \beta_1 x_1 + \beta_* (x_2 + x_3) + \varepsilon$$

将我们已知的 $\beta_2 = \beta_3$ 整合到模型中去,由此便解决了共线性的问题(同时也使得检验 $\beta_2 = \beta_3$ 这一已知信息成为可能)。

尽管几种解决共线性的方法是分开讨论的,但它们仍有许多共同之处:

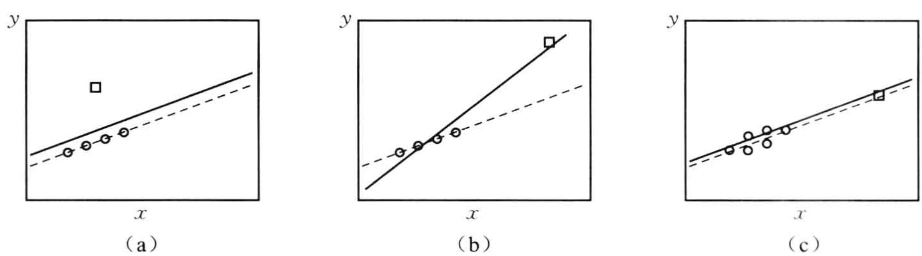
第一,模型的重新确认涉及变量的选择,而变量的选择也有效地重新确认了模型。第二,变量选择潜在地使回归因子不能被全部删除。第三,如果被删除的变量是不为 0 的 $\beta$ 值,且与包含在模型中的变量相关,变量选择就会导致有偏的系数估计值。第四,某些类型的预知信息将导致一个重新确认的模型(正如在假设的例子所示)。第五,可以证明,类似于岭回归的有偏估计方法潜在地对 $\beta$ 的值进行了预先约束。

从这些比较中,我们首先可以得到的经验是,机械的模型选择让与修改程序这一做法将掩饰建模决定的很多推论。因此,这些方法通常不能弥补数据的缺点,也不能作为判断和思考的替代物。

# 第 4 章 | 奇异值与强影响数据

不寻常的数据在最小二乘回归中往往是有问题的，因为他们将严重影响分析结果，并且它们的存在往往表明现有的回归模型不能很好地捕捉到数据的重要特点。一些重要的区别在描述简单回归模型 $y=\beta_0+\beta_1x+\varepsilon$ 的图 4.1 中可以看到。

在简单回归中，一个奇异值是指因变量的值在给定自变量的值时，得到不寻常的观测值。相比，一个单变量的奇异值是 $y$ 或 $x$ 在无条件的情况下，得到不寻常的值，而这样的值不一定是回归中的奇异值。回归中的奇异值在图 4.1 的(a)与(b)中均出现。在图 4.1(a)中，奇异的观测量在 $x$ 的分布中处于中央，因此若删除这个奇异值，对于最小二乘回归的斜率 $b_1$ 与截距 $b_0$ 几乎没有影响。在图 4.1(b)中，奇异值在 $x$ 值上有不寻常的值，因此若将其删除，将显著影响斜率与截距。因为有不寻常的 $x$ 值，所以在图 4.1(b)中，最后一个观察值对回归系数有强烈的影响，但是图 4.1(a)中的中间观测值则为一个弱影响点。




注：(a)一个接近 $x$ 均值的奇异值对回归系数没有很大的影响。(b)一个与 $x$ 均值相距较远的奇异值对回归系数产生了较大影响。(c)一个与其他数据在一条直线上的高影响力观测值并不影响回归系数。

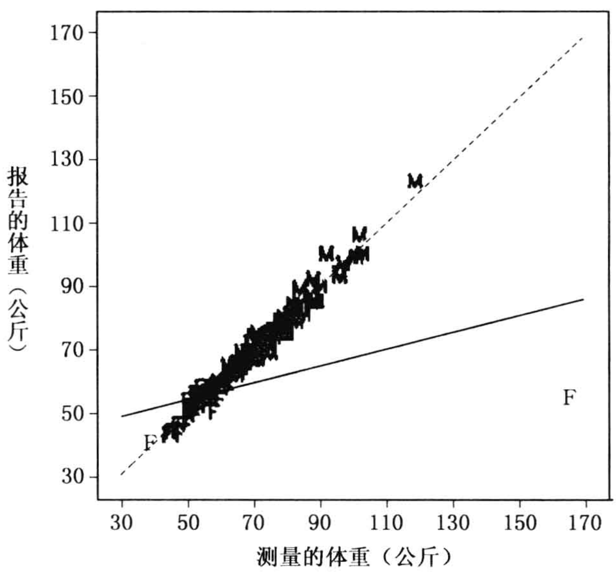
**图 4.1 简单回归分析中的影响力与影响程度**

高影响点组成的奇异值对回归系数有极大的影响。在图 4.1(c)中，最后一

个观测值对回归系数没有影响，哪怕它是一个高影响点，原因在于这个观测值并没有离开剩余的数据组成的直线。下面这个公式将帮助区分这些概念：

$$对系数的影响 = 影响力 \times 差异程度$$

图 4.2 是来自 Davis(1900)真实数据的一个简单而明显的例子。这些数据记录、测量并报告了 183 位参与这一日常生理测试项目的男性与女性的体重(公斤)。作为一项大型研究的一部分，研究者对于判断研究对象是否准确报告他们的体重值以及男性和女性在报告体重方面是否有差异十分感兴趣(这项公开研究仅基于女性研究对象以及后来加入的、从不运动的女性的数据)。Davis(1900)提供了测量的与报告的体重值之间的相关系数。




注：分析中的 183 个对象都参加日常运动。实线显示的是对女性的最小二乘回归，折线显示的是男性的回归。

资料来源：C. Davis。

**图 4.2 对以公斤计的汇报体重根据测量体重和性别进行的回归**

根据申报体重($RW$)和测量体重($MW$)进行最小二乘回归，用一个虚拟变量表示性别($F$:1 为女性，0 为男性)，并用一个交互回归因子可以得到下面的结果(包括括号中系数的标准误)：

$$\widehat{RW} = 1.36 + 0.990MW + 40.0F - 0.725MW \times F$$
$$(3.28) \quad (0.043) \quad (3.9) \qquad (0.056)$$
$$R^2 = 0.89 \quad s = 4.66$$

如果这些结果是严谨地计算出来的，我们则可以总结男性较之女性，更准确地汇报了他们的体重(因为 $b_0 \approx 0$ 且 $b_1 \approx 1$)。另外，体重较轻的女性倾向于报高体重，体重较重的女性倾向于报低体重。但是，图 4.2 显示，男性和女性的不同测试结果实则因一个报告自己的体重为平均值，但测量体重极大的女性研究对象造成的。

事实上，这个研究对象的测量体重和其身高(厘米)在数据输入的时候被搞混了，Davis 在计算出类似的报告与测量体重间的较低的相关系数后发现了这一点。修正数据后，得到如下回归结果：

$$\widehat{RW} = 1.36 + 0.990MW + 1.98F - 0.0567MW \times F$$
$$(1.58) \quad (0.021) \quad (2.45) \quad (0.0385)$$
$$R^2 = 0.97 \quad s = 2.24$$

结果显示，男性与女性均较准确地报告了他们的体重。

还有另外一种方法可以分析 Davis 的体重数据：某一研究者的兴趣在于确定研究对象是否足够准确地报告了他们的体重，以便利用报告的体重作为测量体重的替代数据。因为这种做法会使收集体重数据的支持大大降低，我们自然会认为报告体重受到“真实”体重的影响，正如上面的回归中将报告体重作为因变量。然而替代的问题则基于对测量体重根据报告体重进行回归，下面的回归结果是基于未修正的数据：

$$\widehat{MW} = 1.79 + 0.969W + 2.07F - 0.00953MW \times F$$
$$(5.92) \quad (0.076) \quad (9.30) \quad (0.147)$$
$$R^2 = 0.70 \quad s = 8.45$$

请注意，在此方程中，奇异值对回归系数没有什么影响，原因在于这个奇异值的报告体重值接近女性报告体重值的均值。然而，它对于复相关系数和标准误仍有显著的影响。对于修正后的数据，$R^2 = 0.97$，$s = 2.25$。

## 测量影响力：预测值

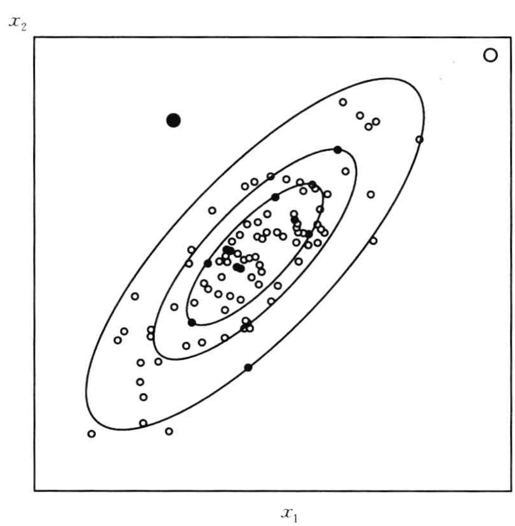
预测值 $h_i$ 是对回归影响较大的一个常见的测量。如此命名这些值的原因是可以通过观测值 $y_i$ 来表达拟合值 $\hat{y}_j$：

$$\hat{y}_j = h_{1j}y_1 + h_{2j}y_2 + \cdots + h_{jj}y_j + \cdots + h_{nj}y_n = \sum_{i=1}^{n} h_{ij}y_i$$

因此，权重 $h_{ij}$ 显示了 $y_i$ 对 $\hat{y}_j$ 的影响程度：如果 $\hat{y}_j$ 很大，则第 $i$ 个观测值对于第 $j$ 个拟合值有较大的影响力。方程也可以写做 $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$，则奇异值 $h_i = h_{ii}$ 表示了 $y_i$ 对所有拟合值的潜在影响。预测值的取值在 $1/n$ 与 1 之间（即 $1/n \leqslant h_i \leqslant 1$），其平均值为 $\bar{h} = (k+1)/n$ 。

在简单一元回归分析中，预测值测量了距离 $x$ 均值的距离：

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$




注：有两个高影响力的点：一个（用较大的空心点标示）在 $x_1$ 与 $x_2$ 上都有异常大的值；另一个（用较大的实心点标示）只是 $x_1$ 与 $x_2$ 的组成较为异常。

**图 4.3 自变量为 $k=2$ 的恒定影响力（恒定的 $h_i$）的等高线**

在多元回归中，$h_i$ 测量了距离 $x$ 圆心的距离，在考虑到 $x$ 的相关结构后，可以由图 4.3 来表示 $k=2$ 时的情况。在 $x$ 空间中多变量的奇异值为强影响的观测值。

在 Davis 根据测量体重对报告体重进行的回归中，最大的预测值是第 12 个观测对象，其测量体重被错误地记录为 166 公斤，即 $h_{12}=0.714$。这个数字远远大于平均的奇异值 0.0219($\bar{h}=(3+1)/183=0.0219$)。

## 查找奇异值：学生残差

为了确认一个奇异的观测值，我们需要一个指标测量 $y$ 在给定 $x$ 下不寻常的程度。一般来说，差异性的观测值都有较大的残差，但是即便误差 $\varepsilon_i$ 有相同的方差(正如回归模型中假设的)，残差也并不相同，$V(e_i)=\sigma^2(1-h_i)$。因此，强影响的观测值往往有较小的残差，这是可以理解的，因为这个观测值可以强迫回归平面向它们靠近。

尽管我们可以通过计算 $e_i'=e_i/s\sqrt{1-h_i}$ 得出标准化残差的值，但这个方程的缺陷在于分子与分母并非独立的，使得 $e_i'$ 无法服从 $t$ 分布。当 $|e_i|$ 很大时，由于包含了 $e_i^2$，$s=\sqrt{\sum e_i^2/(n-k-1)}$ 也同样很大。然而假设删除了第 $i$ 个观测之后重新拟合回归模型，基于剩下的数据得到了一个 $\sigma$ 估计量 $s_{(-i)}$，此时的学生残差为：

$$t_i=\frac{e_i}{s_{(-i)}\sqrt{1-h_i}} \qquad [4.1]$$

该方程有独立的分子与分母，且服从自由度为 $n-k-2$ 的 $t$ 分布。

另一种用于寻找学生残差的步骤使用了均值漂移的模型：

$$y_j=\beta_0+\beta_1x_{1j}+\cdots+\beta_kx_{kj}+\gamma d_j+\varepsilon_j \qquad [4.2]$$

其中 $d$ 是一个虚拟变量集，1 为第 $i$ 个观测值，0 为其他所有观测值。因此有：

$$E(y_j)=\beta_0+\beta_1x_{1j}+\cdots+\beta_kx_{kj}\,(j\neq 1)$$

$$E(y_j) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \gamma$$

如果在检验数据之前,我们就已经怀疑观测 $i$ 与其他的不同,那么自然需要确立方程 4.2。然后为了检验 $H_0: \gamma = 0$,我们可以得到 $t_i = \hat{\gamma}/\text{SE}(\hat{\gamma})$,在 $H_0$ 假设下,其分布为 $t_{n-k-2}$。通过观察则可以发现,它即方程 4.1 中的学生残差。

这与统计中其他方面一样,术语并非完全准确:$t_i$ 有时候也称为"删除学生残差"、"外部学生残差"或者"标准化残差"。最后一个称谓也常应用于$e_i'$。因此,精准地确认电脑程序究竟计算的是哪一个量是非常重要的,但在大样本中,往往有 $t_i \approx e_i' \approx e_i/s$。

因为在大多数的应用中,我们都不会提前怀疑某个观测值,因此我们可以重新拟合均值漂移模型 $n$ 次,且每次针对一个观测值,得到 $t_1, t_2, \cdots, t_n$。在实践中,方程 4.1 与方程 4.2 的其他替代方程对于 $t_i$ 几乎没有计算的效用。因而我们的兴趣往往转向了最大的绝对值 $t_i$,称为 $t^*$。因为我们已经获得了 $n$ 个检验中最大的统计量,所以仅仅去发现 $t^*$ 的统计显著性就不正确了。例如,即便我们的模型是足够的,且暂时不考虑 $t_i$ 之间的相关性,我们仍可以预计观测到 5%的 $t_i$ 在 $t_{0.025} \approx \pm 2$ 之外,1%的 $t_i$ 在 $t_{0.005} \approx \pm 2.6$ 之外等等。

解决同时推论这一问题的办法之一是对最大的 $t$ 值进行 Bonferroni 转换来转化 $t_{n-k-2}$(另一种方法是考虑到学生残差的数量,构造一个分位数比较的散点图,这将在第 5 章中介绍)。Bonferroni 检验需要一个特殊的 $t$ 表格或一个对处于分布尾处的 $t$ 可以获得准确的 $p$ 值的计算机程序。在后一种情况下,假设 $p' = \Pr(t_{n-k-2} > |t^*|)$,则检验 $t^*$ 统计显著性的 $p$ 值为 $p = 2np'$。其中,2 表示这个检验是双尾的,即我们既想要找到大的负奇异值,也想找到大的正奇异值。方程中的 $n$ 表示同时进行了 $n$ 个检验,暗示要在 $n$ 个检验统计量中选择最大的。Beckman Cook(1983)证明了 Bonferroni 转换非常适用于检验最大的学生残差。需要注意的是,要想获得统计的显著结果,则与普通的个别 $t$ 检验相比,需要获得一个更大的 $t^*$。

在 Davis 根据测量体重对报告体重进行的回归中,最大的学生残差是第 12 个观测值:$t_{12} = -24.3$。在这里,$n - k - 2 = 183 - 3 - 2 = 178$,且 $\Pr(t_{178} >$

24.3) $\ll 10^{-8}$ (符号≪表示“远远小于”)。我用于寻找尾处概率值的计算机程序无法计算出一个这样大的 $t$ 的更精确的结果。对于这个奇异值的检验，Bonferroni $p$ 值为 $p \ll 178 \times 2 \times 10^{-8} = 4 \times 10^{-6}$ (即 0.000004)，是一个非常确定的结果。

到目前为止，我已经将确认(以及暗含的潜在修改、去除或调试)奇异值作为一个假设检验。尽管这是目前实际应用中最普遍的一种方式，但还有一个考虑拒绝可能的奇异观测值时，对估计进行投入产出分析的更合理的方法。

假设此时拥有最大 $t_i$ 值的观测是一个不寻常的数据点，但却是通过假设的统计模型计算出的，即 $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$，其中 $\varepsilon_i \sim \text{NID}(0, \sigma^2)$。删除这样一个数据点将会降低估计的有效性，因为模型是正确的(包括正态分布这一假设)，所以最小二乘估计量对 β 的所有无偏估计是最有效的。但是如果这个有问题的数据点与其他的不一致(例如均值漂移中所示的)，则将其删除会使这个估计更加有效。Anscombe(1960)通过做一个保险的类比表明了这一观点：为了获得没有“坏”数据的保障，我们选择了一个拒绝奇异值的规则(或者使用了一个可以抵抗奇异值的估计量，即稳健估计量)，而当这一规则拒绝了“好”数据时，其保险费则由有效性来赔付。

用 $P$ 来代表理想中的保险费，即如果这个模型适用于所有数据，则会导致估计量的均方误上升 5%。用 $z$ 表示相应尾部概率 $P(n-k-1)/n$ 的单位正态变异。根据 Anscombe 和 Tukey(1963)的步骤，计算 $m = 1.4 + 0.85z$，则可以获得：

$$f = m\left[1 - \frac{m^2 - 2}{4(n-k-1)}\right] \times \sqrt{\frac{n-k-1}{n}} \qquad [4.3]$$

$$t' = \frac{f\sqrt{n-k-2}}{\sqrt{n-k-1-f^2}} \qquad [4.4]$$

最后，如果 $|t^*| > t'$，则不具有最大学生残差的观测值。在实际应用中，我们应该探究这些不一致的观测值(将在本章最后讨论)。

例如，在 Davis 起初的 $n = 183$, $k = 3$ 的回归中，$P$ 值为 0.05，则我们得到：

$$P(n-k-1)/n = 0.05(183-3-1)/183 = 0.0489$$

从单位正态表中可以查出，$z = 1.66$，则 $m = 1.4 + 0.85 \times 1.66 = 2.81$。因为 $t^* = 24.3$，远远大于 $t'$，则第 12 个观测值被确认为奇异值。

## 测量影响程度：Cook 距离与其他诊断方法

正如在前面提到的，对回归系数的影响包括影响力和差异程度两部分。最直接的测量影响程度的方法就是逐步删除各个观测值，观察对于系数的影响：

$$d_{ij} = b_j - b_{j(-i)} (i = 1, \cdots, n; j = 0, \cdots, k)$$

其中，$b_{j(-i)}$ 表示当第 $i$ 个观测值被删除时，对 $\beta_j$ 进行最小二乘估计得到的结果。为了便于理解，有必要对 $d_{ij}$ 根据系数标准误的估计值进行度量。

$$d_{ij}^* = \frac{d_{ij}}{\mathrm{SE}_{(-i)}(b_j)}$$

沿用 Belsley 等人的命名(1980)，$d_{ij}$ 往往称为 $\mathrm{DFBETA}_{ij}$，$d_{ij}^*$ 则称为 $\mathrm{DFBETAS}_{ij}$。

使用 $d_{ij}$ 与 $d_{ij}^*$ 的一大问题在于其数量太多，每个都有 $n(k+1)$ 个。当然，运用图形来检验这些值比使用数字的表格省事得多。例如，我们可以对 $d_{ij}^*$ 的每一个系数 $j = 0, 1, \cdots, k$ 构建一个“索引三点图”，只要简单地根据横轴表示 $d_{ij}^*$，纵轴表示每个观测的索引 $i$ 绘制散点图即可。但是，对于这一拟合图上每个观测点的影响程度，则有必要根据其索引进行归纳。

Cook(1977)提出了通过计算“假设” $\beta_j = b_{j(-i)}$，$j = 0, 1, \cdots, k$ 的 $F$ 统计量值，来测量 $b_j$ 与相应的 $b_{j(-i)}$ 之间的“距离”。这个统计量通过对每一个观测 $i=1, \cdots, n$ 进行重新计算获得。获得的值不能直接被当做 $F$ 检验，因为 Cook 的方法只是与检验类似的一种类比，其目的在于获得能够测量独立于 $x$ 度量之外的距离。Cook 统计量可以写为：

$$D_i = \frac{e_1'^2}{k+1} \times \frac{h_i}{1-h_i}$$

其中第一个方程是测量差异程度,第二个是测量影响力(见附录 5)。我们寻找比其他值大的 $D_i$。

Belsley 等人提出了非常相似的测量:

$$\text{DFFITS}_i = t_i\sqrt{\frac{h_i}{1-h_i}}$$

注意,除了非常特殊的数据结果,$D_i \approx \text{DFFITS}_i^2/(k+1)$。另外,还有其他测量影响程度的方法(参见 Chatterjee & Hadi, 1998)。

因为所有删除统计量都基于预测值和残差,我们可以利用图示的方法获得大致的影响程度的测量,即根据 $t_i$ 绘制 $h_i$ 的散点图,并寻找两者较大的观测值。这一散点图更加合适的版本是展示与 Cook 的 $D$ 成比例的圆圈图形以替代散点(参见 Chatterjee & Hadi, 1988:38)。接下来,我们对拥有最大的 $D_{i'}$、$|\text{DFFITS}_i|$ 或者由大的 $h_i$ 与 $|t_i|$ 构成的组合进行观测,检验其 $d_{ij}$ 或 $d_{ij}^*$。

在 Davis 根据测量体重对报告体重的回归中,所有有影响力的点的指标数对有显著差异的第 12 个观测如下:

$$\text{Cook 的 } D_{12} = 85.9(\text{第二大的为 } D_{21} = 0.065)$$

$$\text{DFFITS}_{12} = -38.4(\text{第二大的为 DFFITS}_{50} = 0.512)$$

$$\text{DFBETAS}_{0,12} = \text{DFBETAS}_{1,12} = 0,\ \text{DFBETAS}_{2,12} = 20.0,$$

$$\text{DFBETAS}_{3,12} = -24.8$$

需要注意的是,第 12 个观测值是以女性观测为对象的,对男性样本的截距 $b_0$ 和斜率 $b_1$ 没有影响。

在发展对回归的影响程度这个概念的过程中,我聚焦于对回归系数的改变。但是其他的回归结果也同样应该被检验。其中一个重要结果就是系数的方差和协方差,代表估计的精度。例如,在图 4.1(c)中,其中一个高影响力的点并没有对回归系数产生影响,因为它与其他数据处于一条直线上。在简单一元回归中,估计的最小二乘斜率标准误为 $\text{SE}(b_1) = s/\sqrt{\sum(x_i - \bar{x})^2}$,因此通过增加 $x$ 的方差,高影响力观测点会使得 $\text{SE}(b_1)$降低,即使它并不影响 $b_0$ 与 $b_1$。根据不同情况,这一类观测值可以被认为是有益的(提高了估计的精度),也能

够使我们对估计的 $b_1$ 更有信心。

在多元回归中，我们可以检验逐个删除观测值对 β 的联合置信区域大小的影响。回顾第 2 章可以发现，这一区域的大小与单一系数置信区间的长度相似，因此也与系数的标准误成比例。因此，置信区间长度的平方与系数的样本方法也成比例，同样，联合置信区域大小的平方与"广义"的一系列系数的方差成比例。一项由 Belsley 等人提出的测量影响程度的方法与删除的、全部数据的置信区域比例高度近似：

$$\text{COVRATIO}_i = \frac{1}{(1-h_i)\left(\dfrac{n-k-2+t_i^2}{n-k-1}\right)^{k+1}}$$

可替代的相似方法由其他几位学者提出(Chatterjee & Hadi, 1988)。在此我们需要寻找 $\text{COVRATIO}_i$ 与 1 相差较大的值。

正如对回归系数影响程度的测量，预测值与学生残差也被包括在 COVRATIO 之内。较大的预测值将导致较大的 COVRATIO，但是，即便(事实上，也正是当)$t$ 值很小，一个高影响力且与其他数据在一条线上的观测也会提高估计精度。相比之下，一个差异程度较大但影响力较低的观测将不会使系数有什么改变，因为它通过增大估计的误差方差而降低了回归精度。因此，具有较小的 $h$ 值与较大 $t$ 值的观测将产生一个远远小于 1 的 COVRATIO。

例如，在 Davis 最初的回归中，最极端的值为 $\text{COVRATIO}_{12} = 0.0103$。在这个例子中，一个非常大的 $h_{12} = 0.714$ 远非一个很大的 $t_{12} = -24.3$ 可以弥补。

回归分析的其他特征也可能被个别的观测所影响，包括共线性的程度。尽管对共线性影响程度的正规分析远不止在此所讨论的(见 Chatterjee & Hadi, 1988)，但以下的各条建议仍然非常有用：

第一，对共线性的影响是反映在对系数标准误影响中的一个因素。对误差方差的影响和对 $x$ 变异程度的影响，都可以作为 COVRATIO 这类测量的因素。同样，COVRATIO 和其他类似的测量也可以检查抽样方差和所有回归系数的协方差，包括常数项。然而，我们对共线性问题关注的原因在于，它会影响估计

的精度，而 COVRATIO 正可以评估全面的估计精度。

第二，共线性—强影响点是那些诱因或可以使 $x$ 之间相关变弱的点。这些点往往(但并非总是)有较大的预测值。反之，有较大预测值的点也往往会影响共线性。

第三，单个诱导共线性的点显然存在问题。但那些明显减弱共线性的点同样值得注意，因为它们有时使我们对得出的结论过于自信。

最后，通过对一个自变量根据另一个自变量绘制散点图，往往可以得到共线性—强影响的观测。但是如果共线性问题涉及的自变量超过两个，这种方式就会失效。

## 诊断统计量中的数值截断点

在利用测量影响力和影响程度判断值得注意的观测点时，我刻意没有建议某些数值的标准。我认为，检验这些未知量的分布以确定这些奇异观测值的位置更加有效。对于学生方差来说，假设检验与保险化的观点将得出各种各样的截断点，但是这些我们熟知的标准并不能取代图示检验残差的方法。

然而尽管截断点并不十分重要，但它仍有一定的用处，它可以帮助强化图形的展示。一条水平线可以在一个索引散点图上画出，例如，吸引对超过截断点的值的注意力。同样，这样的值在图中也可以单独被确认(如 Chatterjee & Hadi，1988:38)。

一个诊断统计量的截断点可能是统计理论的产物，或者是通过检验这个统计量的样本分布得出的。截断点可以是绝对的，也可根据样本规模进行调整(Belsey et al.，1980)。对于一些诊断统计量，例如对影响程度的测量，绝对的截断点无法在大样本情况下确认需要注意的观测。这一特征部分反映出大样本可以弥补异常数据而无需大幅度地改动结果，但是截断点往往还是可以辨别影响力相对较大的点，哪怕并不存在具有强烈绝对影响的观测。

接下来对截断点的简单叙述是基于对统计理论的应用得出的。另一个非常简单但广泛使用的基于数据的标准，是检验对一个诊断测量来说最极端的

5%的值。

预测值：Belsley 等人(1980)建议，那些超过 $(k+1)/n$ 均值两倍以上的点就需要注意。这种基于样本规模进行调整的截断点，是当 $x$ 呈多元正态分布且 $k$ 与 $n-k-1$ 都相对较大时，通过近似确认最极端的 5%案例而获得的。但是，这种方法只是一种粗略的方针(关于其他预测值截断点的讨论，参见 Chatterjee & Hadi, 1988)。

学生方差：除了考虑上述讨论的"统计显著性"和估计量稳健性与有效性，关注相对较大的残差值也很有帮助。回顾以前的内容我们知道，在理想状况下，学生残差的 5%是在 $|t_i| \leqslant 2$ 范围之外的。因此，在学生残差的图示上画出 $\pm 2$ 这两条线并在这个区间外进行观测是有意义的。

测量影响程度：对于不同的影响程度的测量方法，有多种建议的截断点：

首先是回归系数的标准化变化。将 $d_{ij}^*$ 根据标准误进行量化后，$|d_{ij}^*|>1$ 或 2 的就是绝对的截断点。然而，正如上面所解释的，这个标准在大样本数据中无法确认异常观测值。Belsley 等人推荐将这种基于样本规模调整的截断点 $2/\sqrt{n}$ 作为需要关注的 $d_{ij}^*$。

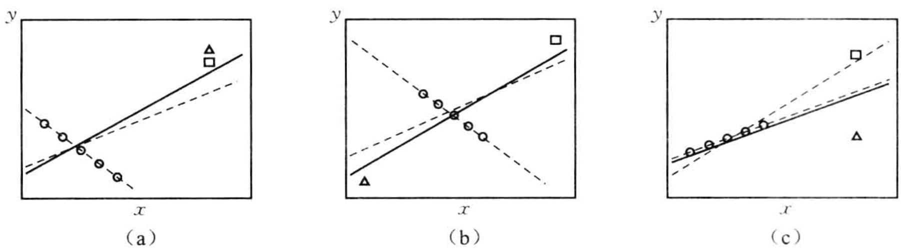
其次是 Cook 的 $D$ 与 DFFITS。对于 Cook 的 $D$ 与 DFFITS，有许多值得推荐的数值截断点，例如考察 $D$ 与 $F$ 统计量之间的类同处。Chatterjee 和 Hadi(1988) 建议比较 $|\text{DFFITS}_i|$ 与基于样本规模的截断点 $2\sqrt{(k+1)/(n-k-1)}$ (参见 Cook, 1977; Belsley et al., 1980; Velleman & Welsch, 1981)。此外，由于 DFFITS 与 Cook 的 $D$ 之间的近似关系，所以这两种测量的截断点的相互转化是很容易的。以 Chatterjee 和 Hadi 的标准为例，我们可得转换后的截断点为 $D_i > 4/(n-k-1)$。绝对的截断点，例如 $D_i>1$，则可能漏掉强影响数据。

第三是 COVRATIO。Belsley 等人建议当 $|\text{COVRATIO}_i - 1|$ 超过基于样本规模调整的截断点 $3(k+1)/n$ 时，就需要注意 $\text{COVTATIO}_i$。

## 联合的强影响观测子集：偏回归图

正如图 4.4 所示，观测子集可以联合造成影响或者改变其他子集的影响。

强影响的子集或者多元奇异值往往可以通过逐一进行观测诊断而确认。但重要的是如何在删除这样的影响点之后重新拟合模型,因为单个强影响值的存在可能强烈影响对其他点的拟合。因此,逐步进行检测的方法并不能保证永远成功。




注:在每个例子中,实线是对所有数据的回归,较细的虚线是将三角点删除后的回归,较粗的虚线是方形点与三角点都删除后的回归。(a)联合强影响观测彼此之间很接近。(b)联合强影响观测分布在数据的两侧。(c)观测彼此之间抵消了:在两个观测点都删除后的回归与根据整个数据进行的回归是相同的。

**图 4.4 联合的强影响数据**

尽管可能存在普适的统计量用以删除包含多个点的子集,但是子集的数量过多(可能有 $n!/[p!(n-p)!]$ 个规模为 $p$)往往使得这种方法不切实际(但可以参见 Chatterjee & Hadi, 1988; Belsley et al., 1980)。另一种替代的方法则是使用图示。

对判断影响程度尤其有用的图示为偏回归散点图,也叫做"偏回归影响力散点图"或者"添加变量散点图"。$y_i^{(1)}$ 表示通过对除了 $x_1$ 外所有的 $x$ 进行最小二乘回归而得到的残差,也就是来自这个模型的残差:

$$y_i = b_0^{(1)} + b_2^{(1)} x_{2i} + \cdots + b_k^{(1)} x_{ki} + y_i^{(1)}$$

同样,$x_i^{(1)}$ 是对其他 $x$ 进行最小二乘回归得出的残差:

$$x_{1i} = c_0^{(1)} + c_2^{(1)} x_{2i} + \cdots + c_k^{(1)} x_{ki} + x_i^{(1)}$$

这里的符号强调了对残差 $y^{(1)}$ 与 $x^{(1)}$ 的解释,它们是 $y$ 与 $x_1$ 在 $x_2, \cdots, x_k$ 的效果被移除之后剩下的部分。可以证明对 $y^{(1)}$ 与 $x^{(1)}$ 进行最小二乘回归获得的

斜率与全模型多元回归通过最小二乘获得的斜率 $b_1$ 是一样的，且从这个回归获得的残差与从全模型中获得的残差也是一样的，即 $y_i^{(1)} = b_1 x_1^{(1)} + e_i$。请注意这里没有常数项，因为作为最小二乘的残差值，$y^{(1)}$ 与 $x^{(1)}$ 的均值为 0。

将 $y^{(1)}$ 根据 $x^{(1)}$ 作图，使我们可以检测它们对 $b_1$ 的影响力和影响程度。对其他回归系数(包括 $b_0$)，类似的偏回归图可以被构建为：

将 $y^{(j)}$ 根据 $x^{(j)}$ 绘制散点图(其中 $j = 0, 1, \cdots, k$)

对于 $b_0$，我们对"常数回归因子"$x_0 = 1$ 与 $y$ 根据 $x_1$ 至 $x_k$ 进行回归，在这个回归方程中没有常数项。

图 4.5 是一个解释性的偏回归图。这个例子中的数据来自 Duncan(1961)，他对 1950 年时的 45 种职业的评价声望值($P$，通过评价该职业为"好的"或者"很好"的百分比进行评估)根据收入和教育水平进行回归($I$：男性收入至少为 3500美元的百分比；$E$：男性为高中毕业生的百分比)。这个回归的主要目标是获得对那些没有直接进行声望评分，但是有教育程度与收入水平数据的职业的拟合分数。拟合方程为(括号中为标准误)：

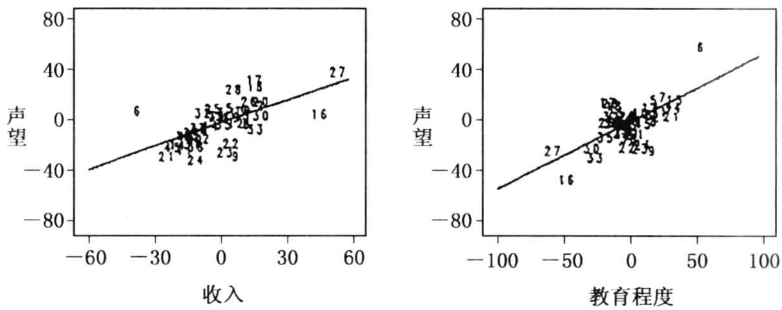
$$\hat{P} = -6.06 + 0.599I + 0.546E$$

$$(4.27) \quad (0.120) \quad (0.098)$$

$$R^2 = 0.83 \quad s = 13.4$$

收入的偏回归图，即图 4.5(a)，显示出 3 个明显会减小回归斜率的强影响观测：(6)部长，他们的工资就其教育程度而言，显得很低；(16)铁路售票员；(27)铁路工程师，他们的收入就其教育程度而言，明显较高。偏回归图中横轴的变量是根据教育对收入进行回归获得的残差，因此，在此方向上偏离 0 的值就是在给定教育水平的情况下收入不正常的观测。

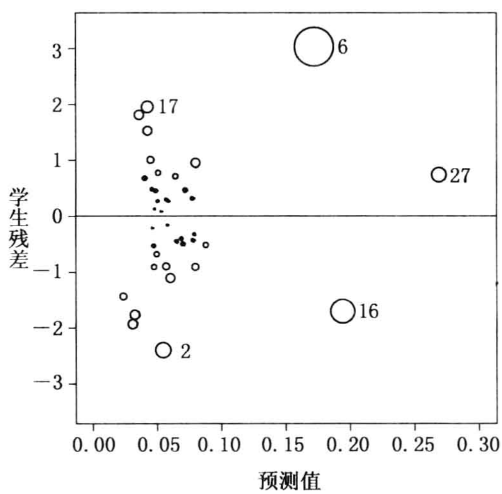
教育的偏回归图，即图 4.5(b)，显示出同样的 3 个对教育系数有相对较高影响力的观测：观测 6 与 16 趋于增大 $b_2$，观测 27 则与其他的数据相隔太远。




注:1950 年 45 个美国职业。每个点的观测号码都在图中显示。如果这个图可以放大,例如在电脑屏幕上,则职业的名称可以显示在数字的旁边。对常数项的偏回归散点图并没有显示于此。

**图 4.5　在对声望根据收入和教育的回归中收入和教育的偏回归散点图**

通过检验单一观测值删除的诊断可以发现,观测 6 具有最大的 Cook 的 $D(D_6 = 0.566)$ 与学生残差($t_6 = 3.14$)。这个学生残差并不非常大,但是,对这个奇异值检验的 Bonferroni $p$ 值为 $\Pr(t_{41} > 3.14) \times 2 \times 45 = 0.14$。图 4.6 是学生残差预测值的散点图,其中圆圈的大小与 Cook 的 $D$ 的值是成比例的。对 $|t_i| > 2$ 或 $h_i > 2(k+1)/n = 2(2+1)/45 = 0.13$ 观测的指标也显示在图中。




注:每个点都围成一个圆形,其面积与 Cook 的 $D$ 是成比例的。当 $h_i > 2h = 0.13$ 或 $|t_i| > 2$ 时,显示出观测的号码。

**图 4.6　对职业声望根据教育和收入的回归中将学生残差根据预测值绘制的散点图**

删除了观测 6 与观测 16 后可获得拟合回归：

$$\hat{P} = -6.41 + 0.867I + 0.332E$$
$$(3.65) \quad (0.122) \quad (0.099)$$
$$R^2 = 0.88 \quad s = 11.4$$

正如在偏回归图中看到的那样，与原来的回归相比，上述方程具有较大的收入斜率和较小的教育斜率。估计的标准误已趋最优，因为相对的奇异值已经被删除了。删除了观测 27，进一步增大了收入的斜率和减小了教育的斜率，但是变化很微小：$b_I = 0.931$，$b_E = 0.285$。

## 非同寻常的数据应该被抛弃吗?

本部分针对将奇异和强影响数据简单抛弃的情况开展讨论。尽管有问题的数据不应该被忽略，但它们也不该被自动和不经考虑地删除。

首先，考虑研究数据为什么非同寻常是非常重要的。真正的坏数据往往可以被修改，如果不能修改，就将之删除。但是，如果一个不一致的数据点是正确的，我们则应该试图去理解为什么这个观测是非同寻常的。例如在 Duncan 的回归中，部长的职业声望很高并不是因为其收入与职业的教育水平。与之类似，我认为相对于其教育水平和职业声望，铁路工人的高收入反映了铁路协会在 20 世纪 50 年代的势力。在这种情况下，我们可能选择对这些奇异的观测分别进行处理。

此外，奇异值或强影响数据可能促成模型的重新确认。例如，奇异数据的模式也许预示着应该引入额外的自变量。如果在 Duncan 的回归中，我们可以确认一个造成部长这一职业非同寻常的高职业声望的因素，且我们可以对其他职业进行这一因素的测量，则这个因素可被引入回归。在一些情况下，对因变量或某一自变量的转化可能使误差的分布对称或者消除共线性，也使奇异值向其余数据靠近。然而，我们必须谨慎以避免过度拟合数据，这会导致一小部分数据决定了整个模型的形式。我们将在第 9 章与第 10 章中重新讨论这一问题。

最后，除非在情况已经非常清楚时，否则我们应该谨慎地删除观测值或者重新确认模型以适应非同寻常的数据。一些研究者合理地采用其他的估计策略，例如稳健回归，它将奇异数据进行较低的权重而非简单地包括或者抛弃它们。这种方法被称为“稳健”的原因在于，即便误差不呈正态分布，它们也可以顺利进行（参见附录 7 中对 Lowess 方法的讨论）。稳健估计的好处通过 Anscombe对保险的类比可以理解：稳健方法与最小二乘方法在误差是正态分布的情况下几乎是同样有效的，并且在奇异值存在的情况下更加有效。这种方法对高度不一致的数据赋予 0 或非常小的权重，但是结果往往与谨慎地应用最小二乘法没有差别，并且事实上，稳健回归的权重也可以用于发现奇异值。此外，大多数稳健回归方法对于高影响力的点非常敏感（参见 Rousseeuw & Leroy，1987）。

# 第 5 章 | 非正态分布误差

误差正态分布这一假设往往带有随意性。然而,中心极限定律确保在非常宽泛的情况下(除了小样本外),基于最小二乘估计的推论往往都是有效的。那么我们应该关注非正态分布的误差吗?

首先,尽管最小二乘估计的有效性是稳健的(正如所讲过的那样,在大样本的情况下,哪怕违背正态分布假定,检验与置信区间的水平也近乎正确),但是这种方法在有效性方面并不稳健:当误差为正态分布时,最小二乘估计量是无偏估计量中最有效的。然而对某些类型的误差分布,尤其在分布具有重尾的情况下,最小二乘估计的有效性将大大降低。在这种情况下,最小二乘估计量则不如其他的替代估计量有效(例如稳健估计量或者被诊断加强的最小二乘法)。在很大程度上,重尾的误差分布是有问题的,因为它们往往导致奇异值,也就是我在前一章所讨论的问题。

常常被引用对最小二乘估计进行辩护的高斯-马尔科夫定理证明,对于观测 $y_i$ 的线性函数,最小二乘系数是最有效的无偏估计量。这一结果基于线性、误差方差一致性以及独立性这些假设,但是并不要求正态分布(参见 Fox,1984:42—43)。尽管这些对线性估计量的限定将导致假定拥有简单的样本属性,但是并不能消除最小二乘法对重尾误差分布的敏感性。

其次,那些高度偏态分布的误差,它们除了容易在偏斜的方向导致奇异值的出现,也容易危及对最小二乘拟合的解释。总之,这种拟合是一种条件均值(在给定 $x$ 下的 $y$),而均值对于一个高度有偏的分布来说也并不是一个对其中心的良好测量。因此,我们倾向于转换数据以获得一个对称的误差分布。

最后,一个多峰的误差分布暗示我们可能忽略了一个或更多可以将数据自然

分成各组的定量变量。由此，对于残差分布的检验则可能引发模型的重新确认。

尽管对非正态误差有检验的方法，在此我仍应该描述一些替代图形的方法来检验残差的分布（参见第 9 章）。这些方法对于认定一个问题的特征以及选择解决的方法更有帮助。

## 残差的正态分位数比较散点图

分位数比较散点图是其中一种图示法，它使我们可以从视觉上比较一个独立随机样本的累积分布（学生残差）与一个累积的参照分布（单位正态分布）。需要注意的是，这里暗示了一种近似，因为学生残差是 $t$ 分布且非独立的，但是这种扭曲往往是可忽略的，至少在中等规模到大规模的样本中是如此。

想要构造分位数比较散点图须满足下列几点：

第一，使学生残差升序排列：$t_{(1)}$，$t_{(2)}$，…，$t_{(n)}$。按照惯例，第 $i$ 个学生残差 $t_{(i)}$ 具有 $g_i = (i-1/2)/n$ 比例的数据在其下方。这种惯例是计算每个观测值以下的一半和其以上一半的值，避免了 0 或 1 部分的累积。0 或 1 部分的累积比例将产生问题，因为我们想要用残差分布去进行比较的正态分布，将永远不会如此接近 0 或 1 的累计概率。

第二，寻找对应 $g_i$ 累计概率的单位正态分布的分位数，也就是 $Z \sim N(0, 1)$ 中满足 $\Pr(Z < z_i) = g_i$ 的 $z_i$ 值。

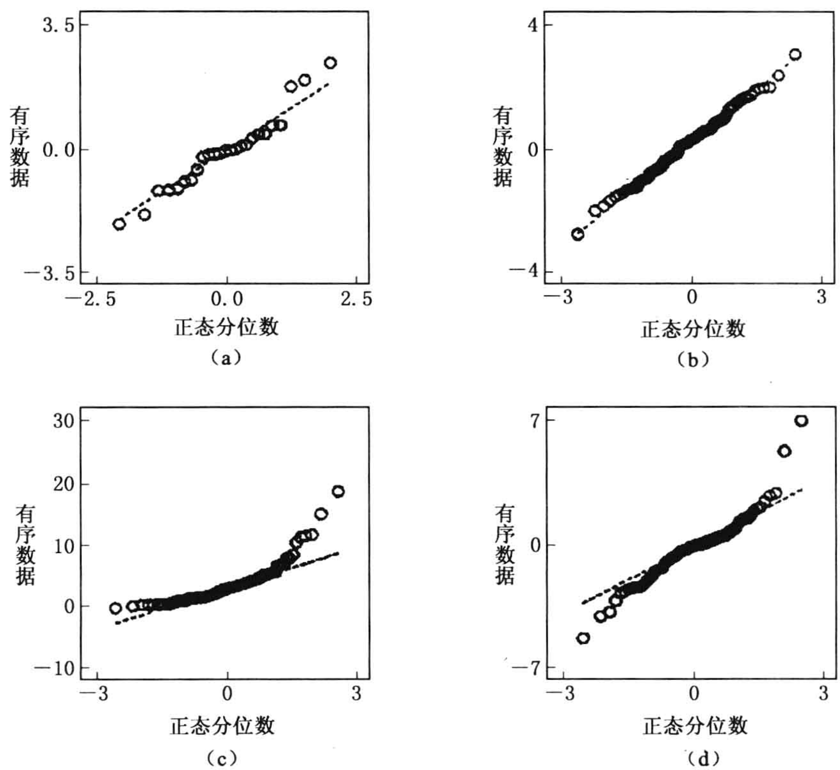
第三，将 $t_{(i)}$ 根据 $z_i$ 绘制散点图。

如果 $t_i$ 是从单位正态分布中获得的，那么在样本误差的界限内，$t_{(i)} = z_i$。因此，我们期望寻找一条拥有为 0 的截距和单位斜率的近似线性散点图，并且有一条直线可以在图中进行比较。相比之下，这个图显示的非线性则可作为非正态分布的症状。

有时将拟合直线根据观测到的中心和残差的散布程度进行调整的做法十分有效。为了理解这种调整是如何达成的，一般假设一个变量 $X$ 是正态分布的，并具有均值 $\mu$ 和方差 $\zeta^2$。那么对于一个值已进行排序的样本，大约有 $x_{(i)} = \mu + \zeta z_i$，其中 $z_i$ 与之前的定义一样。在应用过程中，我们需要估计 $\mu$ 和 $\zeta$，最好利用稳健的方法，因为普通的估计量——样本均值与标准差——会极大地受到极端值的影

响。一般来说，有效地选择是利用 $x$ 的中位数去估计 $\mu$，用 $(Q_3-Q_1)/1.349$ 来估计 $\zeta$，其中 $Q_1$ 与 $Q_3$ 分别是 $x$ 的第一与第三分位数，而中位数和分位数并不受奇异值的影响。需注意的是，1.349 是分离正态分布分位数的标准方差数量。对于学生残差的应用，我们得到拟合直线 $\hat{t}_{(i)}=median(t)+\{[Q_3(t)-Q_1(t)/1.349]\}\times z_i$。本书中的正态分位数比较散点图采用的是最普遍的步骤。

图 5.1 显示了一些对仿真数据的解释性正态概率散点图。图 5.1(a)与图 5.1(b)，其样本规模 $n=25$ 与 $n=100$ 的独立样本是分别从单位正态分布中得到的。图 5.1(c)与图 5.1(d)，其样本规模 $n=100$ 的样本是从高度正偏斜的 $\chi_4^2$




注：(a)来自 $N(0, 1)$，其样本规模为 $n=25$。(b)来自 $N(0, 1)$，其样本规模为 $n=100$。(c)来自正偏态的 $\chi_4^2$，其样本规模为 $n=100$。(d)来自重尾的 $t_2$，其样本规模为 $n=100$。

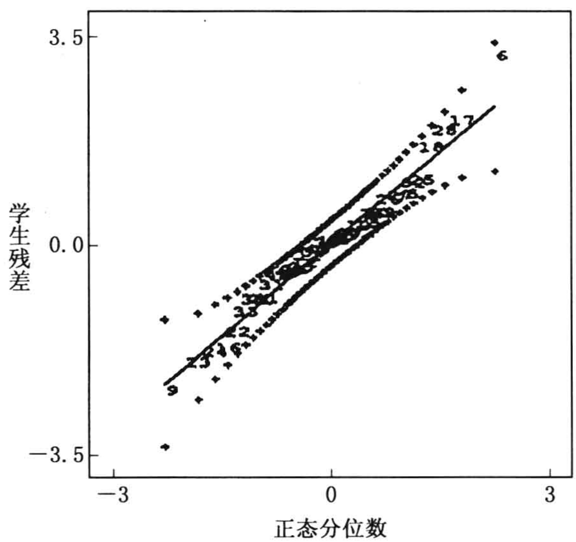
**图 5.1　解释性的正态分位数比较散点图**

分布和重尾的 $t_2$ 分布中分别得出的。请注意图中偏斜与重尾是如何从正态分位数比较散点图中显示出对线性的偏离的。奇异值与相应的正态分位数相比是那些异常大或小的值,由此能很好地进行辨识。

对正态的偏离的判断可以通过对抽样方差的信息来绘制散点图而获得。如果学生残差是从单位正态分布中独立得出的,那么:

$$\mathrm{SE}(t_{(i)}) = \frac{1}{\varphi(z_i)}\sqrt{\frac{g_i(1-g_i)}{n}}$$

其中 $\varphi(z_i)$是单位正态分布的概率密度(即"高度")。因此,在分位数比较散点图中,计算 $z_i \pm 2\times \mathrm{SE}(t_{(i)})$ 可得出在拟合直线 $\hat{t}_{(i)} = z_i$ 附近约 95%的置信区间。如果拟合直线的斜率取值为 $\hat{\zeta} = (Q_3 - Q_1)/1.349$ 而不是 1,估计的标准误则可能需要乘以 $\hat{\zeta}$。而 Atkinson(1985)提出了另一种计算标准误的方法,他建议使用一种不将学生残差视为独立和正态分布的计算性仿真过程。




注:图中显示了一条基于 $t$ 的中位数和分位点的拟合线,以及两条近似±2SE 的逼近线。

**图 5.2 由职业声望根据收入和教育回归获得的学生残差绘制的正态分位数比较散点图**

在直线和两条标准误的界限的问题上,需要注意的是残差的分布是十分合理的。

## 残差的直方图

正态分位数比较散点图的一个优点是在分布的尾部仍有很高的分辨率，这使问题很容易被发觉。但是这种方法的一个缺点在于，它不能呈现出整个残差分布的形状。例如，多元的情况在分位数比较散点图中就难以辨识。

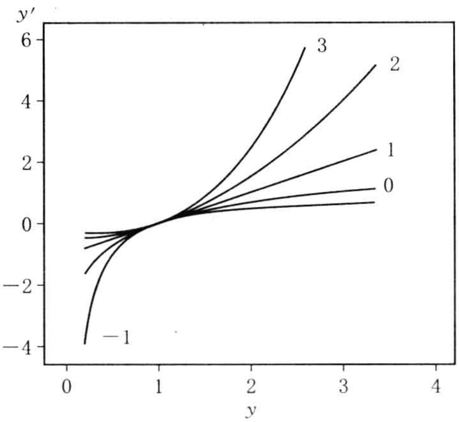
相比之下，直方图（频数柱状图）在分布的尾部或者数据稀少的时候具有较低的分辨率，但是对于呈现一个分布的整体形状十分有效。然而，任意的分类界限、任意的区间以及直方图的不准确性，有时会产生对这一数据的错误印象。这些问题可以通过使直方图变得平滑而部分地解决（参见 Silverman，1986；Fox，1990）。一般来说，对于小样本（即 $n < 100$），我倾向于选择茎叶图，它可以直接记录直方中数值数据值（Tukey，1977）；对于中型规模的样本（即 $100 \leqslant n \leqslant 1000$），则使用平滑直方图；对于大样本（即 $n > 1000$），则采用最优较窄直方的直方图。

| | |
|---|---|
| -2* | 3 |
| -1. | 977 |
| -1* | 41 |
| -0. | 99865 |
| -0* | 444433110 |
| 0* | 000011133334 |
| 0. | 5577788 |
| 1* | 00 |
| 1. | 68 |
| 2* | 0 |
| 2. | |
| 3* | 1 |

**图 5.3　对职业声望根据收入和教育进行回归获得的学生残差绘制的茎叶图**

图 5.3 是对 Duncan 回归中残差的茎叶图。这个图示并没有显示出什么需要注意的问题，因为只有一个单一的节点，分布合理对称。尽管最大值（3.1）与次大值（2.0）相距略远，但是并没有明显的奇异值。

茎叶图中的每一个数据值都可以被分为两部分：前面的数位组成了茎，后面的数位则组成了叶。其后的数位则被删除了，并不进行四舍五入（截断使在表或列中寻找特定值变容易）。对于学生残差，在小数点上很容易进行这一截断。例如在图 5.4 中的残差：0. 3039→0|3；3. 1345→3|1；−0. 4981→−0|4。需要注意的是，每个数位出现两次，形成宽度为 0. 5 的箱柜。有星号标示的茎（例如 1*）对应 0—4 的叶；有点标示的茎（例如 1.）对应 5—9 的叶（更多关于茎叶图的信息，参见 Velleman & Hoaglin，1981；Fox，1990）。




注：标示“$p$”的代表 $y' = (y^p - 1)/p$，对于 $p = 0$，$y' = \log_e y$。

资料来源：Hoaglin, Mosteller & Tukey (eds.), *Understanding Robust and Exploratory Data Analysis* © 1983, John Wiley and Sons Inc.

**图 5.4 一系列幂与根的转换**

## 通过转换矫正不对称

一种针对回归中许多问题的常用有效方法是将数据进行转换，使它们更加符合线性模型的假设。在本章和其后的章节中，我将会介绍一些转化的方法，使误差呈对称分布、误差方差变得稳定，并使 $x$ 与 $y$ 之间变成线性关系。

在上述每种情况下，我们将使用幂与根的形式，将变量 $y$ 进行替换（在这里暂时使用 $y$，其后我们将同样对 $x$ 进行转换），$y' = y^p$。一般来说，$p = -2$、$-1$、$-1/2$、$1/2$、2 或者 3，但有时我们也使用其他的幂与根形式。需要注意的是，$p = 1$ 意味着没有进行转化。当幂为 0 时，$y^0 = 1$，这意味着 $y$ 的值是不变的。当我们使用 $y' = \log y$ 这一形式时，通常使用 2 或 10 作为对数的底，因为对数不同底的转换只有一个常量因素的差别，我们可以选择便于解释的底。将对数转换的使用作为“零次幂”是合理的，因为 $p$ 越接近 0，$y^p$ 越接近对数形式（正规来说，$\lim_{p\to 0}[(y^p - 1)/p] = \log_e y$，其中 $e \approx 2.718$，为自然对数）。最后，对负的幂数来说，我们有 $y' = -y^p$，其值与同样级数的 $y$ 值相反。

当我们愈加偏离 $p = 1$ 的任意方向，这种转化愈加强烈，正如在图 5.4 中所

示。这些转换中的某些影响在表 5.1(a)中可见。幂与根的转化"一点一点往上"时(这一词语来自 Tukey，1977)——即接近 $y^2$——对于扩散 $y$ 的较大值与 $y$ 的较小值有不同的效果;转化"一点一点往下"时——即接近 log $y$ 时——产生相反的效果。为了矫正一个正的偏态分布,如表 5.1(b),需要向下转换;为了矫正应用中往往较少见的负的偏态分布,如表 5.1(c),则需要向上转换。

**表 5.1　通过幂转换矫正偏态**

(a) 中间的数字显示了幂转换的效果

| $-1/y$ | $\log_{10}y$ | ← $y$ → | $y^2$ | $y^3$ |
|---|---|---|---|---|
| −1 | 0 | 1 | 1 | 1 |
| }1/2[a] | }0.30 | }1 | }3 | }7 |
| −1/2 | 0.30 | 2 | 4 | 8 |
| }1/6 | }0.18 | }1 | }5 | }19 |
| −1/3 | 0.48 | 3 | 9 | 27 |
| }1/12 | }0.12 | }1 | }7 | }37 |
| −1/4 | 0.60 | 4 | 16 | 64 |
| }1/20 | }0.10 | }1 | }9 | }61 |
| −1/5 | 0.70 | 5 | 25 | 125 |

(b) 向下进行幂转化以矫正正偏态,拉向右尾

| $y$ | → $\log_{10}y$ |
|---|---|
| 1 | 0 |
| }9 | }1 |
| 10 | 1 |
| }90 | }1 |
| 100 | 2 |
| }900 | }1 |
| 1000 | 3 |

(c) 向上进行幂转化以矫正负偏态,拉向左尾

| $y$ | → $y^2$ |
|---|---|
| 1.000 | 1 |
| }0.414 | }1 |
| 1.414 | 2 |
| }0.318 | }1 |
| 1.732 | 3 |
| }0.268 | }1 |
| 2.000 | 4 |

注:a. 行间的数字表示两个相邻数字间的差异。

我假设所有数据值都是正的——一项幂转化中必需的条件以保持阶数的存在。在实践中，通过加上一个很小的常量，负值便可以在进行转换之前被消除，这个小的常量往往被称为这一数据的“出发点”。同样，为了确保幂转换的效果，最大与最小数据值的比例必须足够的大，否则这种转换将与线性的非常接近。但如果比例较小，则可以用一个负的出发点来解决这个问题。

在回归分析中，通过检验残差分布而发现的有偏误差分布，可以通过对因变量的转换进行矫正。尽管有更复杂的方法(参见第 9 章)，但通过反复试错法，便可获得好的转换。

有下界的因变量也会导致正向有偏分布，利用往下幂转换的效果往往非常好。然而当数据值在下限累积时，也称为“截断”或者“删节”(参见 Tobin, 1958)，幂转换便因此失去效果。同样，同时具有上下限的数据，例如比例和百分比，往往需要其他的解决方法。例如，通过 $y' = \log[y/(1-y)]$ 进行 logit 或“log odds”转换往往对比例很有效。

在回归分析中，对变量的转化也引发了有关解释的问题。我将在第 7 章的末尾简短地处理这些问题。

# 第 6 章 | 不一致的误差方差

## 寻找不一致的误差方差

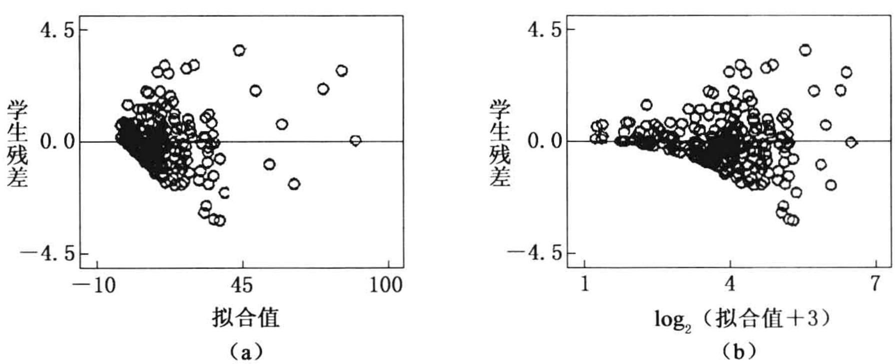
回归模型一直假设因变量的变异在回归平面附近,即误差方差在所有位置是一样的: $V(\varepsilon) = V(y \mid x_1, \cdots x_k) = \sigma^2$。不一致的误差方差往往叫做"异方差性"。尽管在误差方差不一致的情况下,最小二乘估计量仍是无偏且一致的,但它的有效性会受到影响,而系数标准误的普通公式也是不准确的,而问题的严重性取决于误差方差不一致的程度。在本章中,我将叙述一些图形的方法以发现不一致的误差方差这一问题。对异方差性的检验将在第 8 章对离散数据的讨论和第 9 章最大似然估计法中涉及。

因为回归平面是 $k$ 维的,并嵌入在一个 $k+1$ 的空间中,所以一般来说,若 $k$ 大于 1 或 2,就很难直接利用图示检验的方法评估误差方差的一致程度。然而,误差方差往往随着 $y$ 的期望值的增加而增加,或者误差方差和某一 $x$ 之间可能存在系统性的关系。前一种情况可以通过对残差根据拟合值绘制散点图而发现,后一种情况则需要对残差根据每个 $x$ 绘制散点图。需要注意的是,对残差根据 $y$(与 $\hat{y}$ 相对的)绘制散点图往往比较困难。图形可能因此被扭曲:在 $y$ 与 $e$ 之间具有嵌入的相关性,因为 $y = \hat{y} + e$。事实上,$y$ 与 $e$ 之间的相关系数为 $r(y, e) = \sqrt{1 - R^2}$。相比之下,最小二乘拟合确保 $r(\hat{y}, e) = 0$,从而获得一个更容易检查不一致分布的证据的散点图。

因为即便误差具有一致的方差,最小二乘残差仍然具有不等的方差,因此我建议对学生残差根据拟合值绘制散点图。最后,一个变化的分布模式往往在对 $|t_i|$ 或 $t_i^2$ 根据 $\hat{y}$ 绘制的散点图中更容易被分辨出来,并可能被 lowess 平滑散

点图所增强(参见附录 7);当样本规模非常大或者 $\hat{y}$ 的分布非常不均匀时,使这个散点图变得平滑就十分有用。图 6.2 就是一个例子。

图 6.1(a)是对学生残差对拟合值的解释性散点图。在图 6.1(b)中,学生残差是根据 $\log_2(3+\hat{y})$ 绘制的散点图。通过矫正 $\hat{y}$ 中正向偏斜的值,第二个散点图使分辨出残差的分布随着 $\hat{y}$ 的增加而扩散这一趋势变得容易。这个例子的数据来自 Ornstein(1976)对加拿大 248 个大型企业连锁董事会的研究。将每个公司的连锁董事会和行政性职位数量根据公司资产进行平方根转换以使关系变为线性(参见第 7 章);代表 10 个行业分类的 9 个虚拟变量,并以重工业为参照类进行回归;3 个代表 4 个国家的虚拟变量,并以加拿大作为参照类进行回归。表6.1中左边的列显示的是回归的结果。需要注意的是,残差的分布随着 $\hat{y}$ 而分散的原因,一部分是由于 $y$ 的下界 0,因为 $e=y-\hat{y}$,对应某一个 $\hat{y}$ 的最小残差为 $e=0-\hat{y}=-\hat{y}$。




注:(a)$t$ 根据 $\hat{y}$。(b)$t$ 根据 $\log_2(3+\hat{y})$。对数转换降低了集合值的偏态,使得增加的残差更易辨别。

**图 6.1 对 Ornstein 连锁董事会回归中的学生残差根据拟合值绘制的散点图**

## 矫正不一致的误差方差

转换往往可以用于矫正误差方差随着因变量变大而增加(有时会减小)的趋势:如果残差随着拟合值分布范围变广而分散,则使 $y$ 的幂与根向下转换;如果残差随着拟合值分布范围变广而紧缩,则使 $y$ 的幂与根向上转换。通过反复

试错法，可以选出适合的转换（选择一个方差稳定化的转换的方法参见第9章）。如果误差方差对某一个 $x$ 是成比例的，或者如果 $V(\varepsilon_i)$ 的模式是已知的某一常数比例，则另一种对 $y$ 的转换就是权重最小二乘估计（WLS）。对于异方差性，也可以矫正最小二乘系数的估计标准误，White（1980）提出一种方法，参见附录9。这种方法的优点在于不需要了解不一致误差方差模式的信息（例如，方差随着 $y$ 或 $x$ 增大而增大）。但是如果异方差问题很严重，矫正后的标准误比一般公式获得的往往要大得多，如此，发现不一致方差的模式并矫正它（通过转换或WLS估计）将帮助获得更有效的估计。在任意情况下，只有当问题非常严重的时候，才会出现错误矫正不一致误差方差的情况，例如误差的方差的分布随着3个或3个以上的因素而变化（例如，误差方差随着10个或以上的因素而变化，参见附录10）。

对 Ornstein 的连锁董事会回归，平方根转换似乎能够矫正残差随着因变量等级的升高而扩散这一依存关系。图6.2是对转换后的数据将 $|t_i|$ 根据 $\hat{y}_i$ 绘制的散点图。表6.1中右边的列是回归的结果。图6.2中 lowess 平滑后的结果显示，学生残差的平均绝对值并未随着拟合值的上升而发生变化。

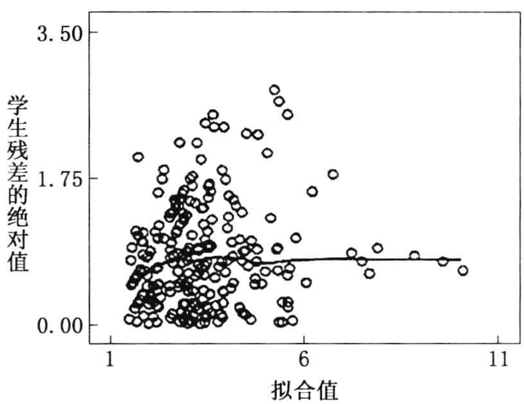
**表6.1 对加拿大284个企业的连锁董事会和行政人员数量根据公司资产、所在部门和州进行的回归**

| 回归因子 | 连锁董事会 | | $\sqrt{连锁董事会+1}$ | |
|---|---|---|---|---|
| | 系　数 | 标准误 | 系　数 | 标准误 |
| 常数项 | 4.19 | 1.85 | 2.33 | 0.231 |
| $\sqrt{资产}$ | 0.252 | 0.019 | 0.0260 | 0.00232 |
| 行业[a] | | | | |
| 　农业、食品 | −1.20 | 2.04 | −0.0567 | 0.255 |
| 　轻工业 | | | | |
| 　冶金 | 0.342 | 2.01 | 0.356 | 0.252 |
| 　木材、造纸 | 5.15 | 2.68 | 0.786 | 0.335 |
| 　建筑 | −5.13 | 4.70 | −0.740 | 0.588 |
| 　运输 | −0.381 | 2.82 | 0.354 | 0.353 |
| 　工商 | −0.867 | 2.63 | 0.148 | 0.329 |
| 　银行 | −14.4 | 5.58 | −2.25 | 0.697 |

续表

| 回归因子 | 连锁董事会 | | $\sqrt{连锁董事会+1}$ | |
|---|---|---|---|---|
| | 系　数 | 标准误 | 系　数 | 标准误 |
| 其他金融 | −5.70 | 2.93 | −0.0880 | 0.366 |
| 控股公司 | −2.43 | 4.01 | −0.245 | 0.502 |
| 控制的国家[b] | | | | |
| 美国 | −8.09 | 1.48 | −1.11 | 0.185 |
| 英国 | −4.44 | 2.65 | −0.527 | 0.331 |
| 其他 | −1.16 | 2.66 | −0.114 | 0.333 |
| $R^2$ | 0.655 | | 0.580 | |

注：a. 虚拟变量的参照类：重工业。
b. 参照类：加拿大。
资料来源：M. Ornstein(个人联系)；这个数据同样被 Fox(1984)所使用。




注：图中的线是 lowess 平滑，使用了 $f = 0.5$ 的二次稳健迭代。

**图 6.2　对学生残差的绝对值根据拟合值绘制的散点图**

表 6.1 中原始与转换后的回归系数并不能直接用来比较，因为因变量的度量被改变了。然而，公司资产仍明显呈现正效应，而控制的国家变量也保有其原先的排列。不同行业间的差别在这两个分析中也相类似，尽管并非完全一致。对两个结果的比较可以发现，这两组虚拟变量的参照类——加拿大和重工业，被默认系数为 0。

将 $y$ 进行转换也使误差分布的形状和对 $y$ 根据 $x$ 回归的形状被改变。通

过转变产生一致的残差变异程度，也往往可以使残差的分布更对称。在某些情况下，消除了不一致的分布也使 $y$ 与 $x$ 之间的关系更接近线性（参见下一章）。然而，这些附产品不一定是矫正误差方差的结果，并且在对 $y$ 进行转换后检查数据的非线性也很重要。当然，由于我们无法在对 $y$ 进行转换之前就知道回归是否是线性的，因此我们应该在对 $y$ 进行转换之前就检验非线性问题。

最后，不一致的残差分布有时是因为忽略了模型中重要影响的证据。假设有一个被忽略的分类自变量（比如地区位置）与公司资产交互影响连锁企业，尤其是公司资产的斜率，即便在每个地区都是正向的，但在某些地区也比其他的地区陡峭。那么即便矫正后的模型中误差有一致的分布，但若忽略了地区和其与公司资产的交互效应，则可能生成一个扇形的残差散点图。因此，为了发现这些特殊的误差，需要深入了解数据产生的过程，而不能简单地依靠诊断。

# 第 7 章 | 非线性

任意处的 $E(\varepsilon)$都等于 0 这一假设暗示着特定的回归平面能够涵盖 $y$ 与 $x$ 之间的依存关系。违背这一线性假设将使模型无法捕捉因变量与自变量之间关系的系统性模式，例如某一被确认为线性的部分关系可能是非线性的，或者两个被确认为具有累加性偏效应的自变量可能是交互影响 $y$ 的。但是，即便回归平面 $E(y)$并没有被准确确认，拟合模型往往也是一种有用的假设。不过在其他情况下，这种模型可能有极大的误导作用。

即便确认了回归因子只是很小数量的基本自变量组成的函数，回归平面也往往是高维的。因此，正如在不一致误差方差的情况下，需要关注偏离线性的特别模式。在本章中，图形的诊断方法用两维的图示代表观测的高维点云 $\{y_i, x_{1i}, \cdots, x_{ki}\}$。利用现代的电脑画图，此处的观点可以扩展至三维，例如对自变量间的二元交互进行相应的检测。

## 残差与偏残差散点图

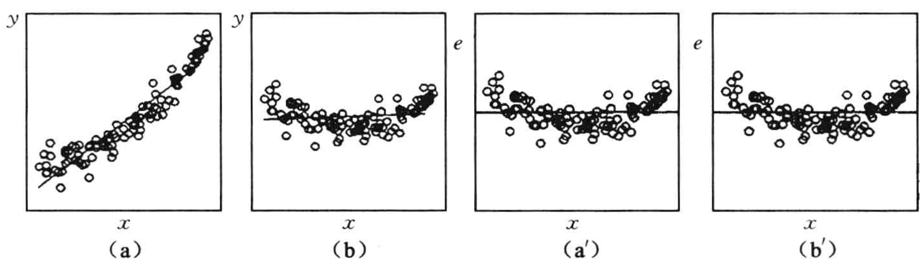
尽管在多元回归中对 $y$ 根据每个 $x$ 绘制散点图是非常有效的，但是这个散点图并不能涵盖全部的情况(有时可能会误导)，因为我们的兴趣聚焦于在控制了所有其他的 $x$ 后，$y$ 与 $x$ 之间的偏相关关系，而不是 $y$ 与单一 $x$ 之间的边际关系。基于残差的散点图在这种情况下则更加适合。

将残差或者学生残差根据每个 $x$ 绘制散点图(可能会被 lowess 平滑而增强，参见附录 7)，对于探寻对线性的偏离非常有效。正如图 7.1 所示，残差散点图不能区分单调(例如严格的增加或减小)和非单调(例如有升有降)的非线性

关系。残差散点图之所以不能捕捉单调与非单调非线性关系之间的区别,原因在于最小二乘拟合确保了残差与每个 $x$ 之间线性的非相关。然而这种区分是非常重要的,正如下面即将讨论的,因为单调的非线性往往可以被简单地转换而矫正。例如在图 7.1 中,案例(a)可以由 $y=\beta_0+\beta_1x^2+\varepsilon$ 建模,而案例(b)无法利用对 $x$ 的幂转换而变成线性的,而需要一个多项式的重新确认来进行处理:

$$y=\beta_0+\beta_1x+\beta_2x^2+\varepsilon$$

然而案例(b)也可以通过调整而对 $x$ 进行转换:$y=\beta_0+\beta_1(x-\alpha)+\varepsilon$,但在此,我不对这种方法进行讨论。




注:残差的散点图并没有分辨出非线性但单调的关系与非线性非单调的关系。

**图 7.1 简单回归的散点图(a)和图(b)与相应的残差散点图(a′)与图(b′)**

与简单的残差散点图相比,后文将介绍的用于寻找强影响数据的偏回归散点图,可以用来揭示非线性并判别关系是不是单调的。这些散点图对于确定一个转换并不一定永远有效,但是偏回归散点图将 $x_j$ 根据其他 $x$ 进行调整,未经调整的 $x_j$ 则被转换以重新确定模型。偏残差散点图,也称做“分量—残差散点图”,往往是另一种有效的方法。但是在检验影响力与影响程度方面,偏残差散点图并不如偏回归散点图有效。

定义第 $j$ 个回归因子的偏残差为:

$$e_i^{(j)}=e_i+b_jx_{ji}$$

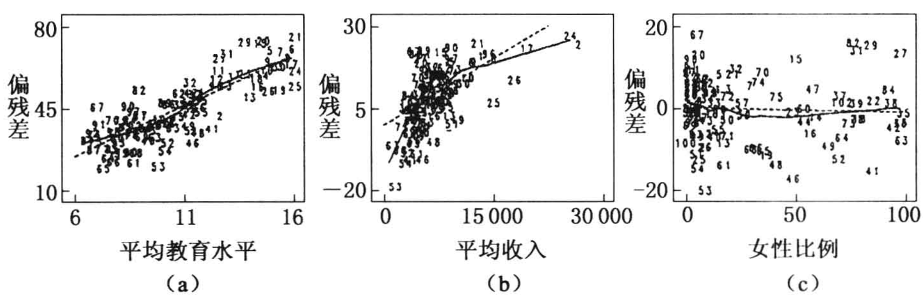
用文字解释为,将 $y$ 与 $x_j$ 的偏相关关系线性分量加回到最小二乘残差上,而这

个残差可能包括未建模的非线性分量。然后将 $e^{(j)}$ 根据 $x_j$ 绘制散点图。通过建构，使多元回归的系数 $b_j$ 成为对 $e^{(j)}$ 根据 $x_j$ 进行简单回归的斜率，但是非线性也同样可以在这个散点图中显现，而 lowess 平滑法也可以帮助解释这个散点图。

图 7.2 中的偏残差散点图是针对将 1971 年 102 个加拿大职业的声望($P$)，根据平均教育年限($E$)和平均收入水平($I$)，以及女性在此职业中的百分比($W$)进行的回归(Pineo & Porter，1967；相关的结果参见 Fox & Suschnigg，1989；Duncan 对类似的美国数据所做的回归)。在每个散点图中都进行了 lowess 平滑。回归的结果如下：

$$\hat{P} = -6.79 + 4.19E + 0.00131I - 0.00891W$$
$$(3.24) \quad (0.39) \quad (0.00028) \quad (0.0304)$$
$$R^2 = 0.80 \quad s = 7.85$$

需要注意的是，回归系数的大小不应被用来进行比较，因为自变量的测量单位不同，尤其是收入的单位非常小(美元)，而教育的单位则相对较大(年限)。我们应该根据每个相应的自变量的单位来解释回归系数。在这个例子中，教育与收入的系数都相对较大，而女性比例的系数则非常小。




注：每个点都附有观测值的索引。在每个图中，都显示了最小二乘拟合(虚线)和 lowess 平滑(实线，$f = 0.5$，进行了二次稳健迭代)。

资料来源：B. Blishen，W. Carroll & C. Moore(个人联系)；加拿大人口普查(加拿大统计年鉴，1971：19.1—19.21)；Pineo & Porter(1967)。

**图 7.2 对 1971 年加拿大 102 个职业的声望评分根据职业特征，以教育、收入和女性比例进行回归获得的偏残差散点图**

教育的偏残差散点图是明显的单调非线性，而收入则更加明显，见图7.2(a)与图7.2(b)。当加入教育与收入水平时，具有中等教育水平的女性的职业具有较低的职业声望这一趋势则没有那么明显，见图7.2(c)。就我而言，在没有经过lowess平滑之前，对教育和女性比例的偏残差散点图的趋势和模式都难以判断，因为对线性的偏离并不严重。收入与女性比例的非线性模式非常简单：对前者而言，lowess曲线是向下开口的；对后者而言，则向上开口。然而对教育而言，弯曲的方向发生了变化，构成了一个更加复杂的非线性模式。

Mallows(1986)指出，偏残差方差的散点图往往能够更清晰地揭示非线性。首先，对模型加入一个多项式使其变为：

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_j x_{ji} + \gamma x_{ji}^2 + \cdots + \beta_k x_{ki} + \varepsilon_i$$

然后，在拟合完模型后，构造"加强"偏残差：

$$e'^{(j)}_i = e_i + b_j x_{ji} + c x_{ji}^2$$

请注意，此处的对 $x_j$ 的回归系数 $b_j$ 与原先模型是不同的，因为原模型并没有包括平方项。最后，对 $e'^{(j)}$ 根据 $x_j$ 绘制散点图。

## 进行线性转换

通过观察图7.3，我们可以考虑幂转换是如何使一个单调的非线性关系线性化的。这里，我们根据 $x = 1, 2, 3, 4, 5$ 对 $y = (1/5)x^2$ 绘制散点图。通过建构让 $y = (1/5)x'$ 中的 $x' = x^2$，则可以使关系变得线性化，或者在 $y' = \sqrt{1/5}x$ 中使 $y' = \sqrt{y}$。图7.3展示了每个转换是如何对其中某一个轴进行不同的伸展，从而使曲线变成一条直线的。

正如图7.4所示，一共有4种简单的单调非线性模式。每一种都可以通过对 $y$、$x$ 或者两者同时在幂与根进行向上和向下转换，曲度的方向决定了在幂与根方向上的移动。Tukey(1977)将这称做"撑压法则"。反复试错法可以帮助找出最合适的线性转换方法。

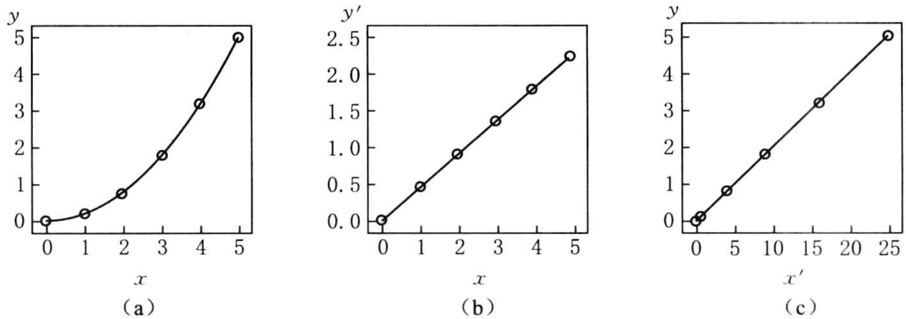
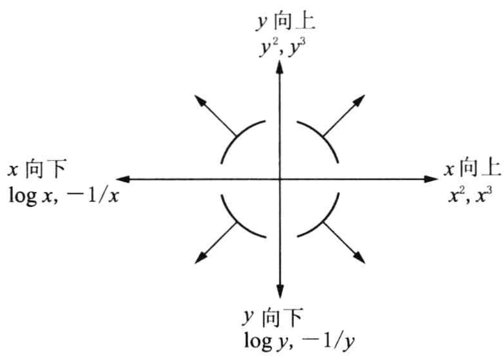



**图 7.3 对 $y$(a 到 b)和 $x$(a 到 c)的转换如何使一个简单的单调非线性关系线性化**




资料来源:Tukey, *Exploratory Data Analysis* © 1977, Addison-Wesley Publishing Co.

**图 7.4 通过“撑压法则”确定一个对线性的转换**

在多元回归中,撑压法则可用于偏残差散点图中。一般来讲,我们倾向于对 $x_j$ 而非 $y$ 进行转换,因为改变对 $y$ 的度量将影响它与其他回归因子之间的关系以及对 $y$ 进行转换将改变误差的分布。而唯一的例外只出现在所有的偏残差散点图中,所有非线性模式都类似的情况下。此外,logit 转换往往在因变量为比例的情况下有效。

与在图 7.1(b)中所显示的一样,非单调非线性(以及一些复杂的单调模式)可以通过对 $x$ 进行多项式拟合来解决,二次方确认在应用中往往非常有效。只要模型对参数来说还是线性的,那么它就可以用最小二乘回归来进行拟合。

对加拿大职业声望数据进行反复试错法可以得出对收入的对数转换。职

业声望与妇女比例之间的曲线偏关系建议我们应当使这个自变量包含线性项与二次项。这些改变对于模型的拟合产生了适度(但可识别)的改进：

$$\hat{P} = -111 + 3.77E + 9.36\log_2 I - 0.139W + 0.00215W^2$$
$$(15) \quad (0.35) \quad (1.30) \quad (0.087) \quad (0.00094)$$
$$R^2 = 0.84 \quad s = 6.95$$

需要注意，妇女比例的二次项具统计显著性。这个变量的偏效应相对较小，但是其范围包括当职业的女性比例为 32%时，最小声望分数的－2.2 到假设的职业的女性比例为 100%时，声望分数的 7.6。因为教育的偏残差散点图的非线性模式是复杂的，对这个因变量的幂转化将不会有效。通过反复试错法，我们知道将教育取平方也只能将 $R^2$ 增加到 0.85。

在对数据进行转换或者重新确定模型的函数形式时，需要对实际情况与建模进行交互的考虑。然而我们必须认识到，社会理论往往并非数学化那样具体，因为理论告诉我们，声望应该随着收入上升，但是它并不能确认这个关系的函数形式。

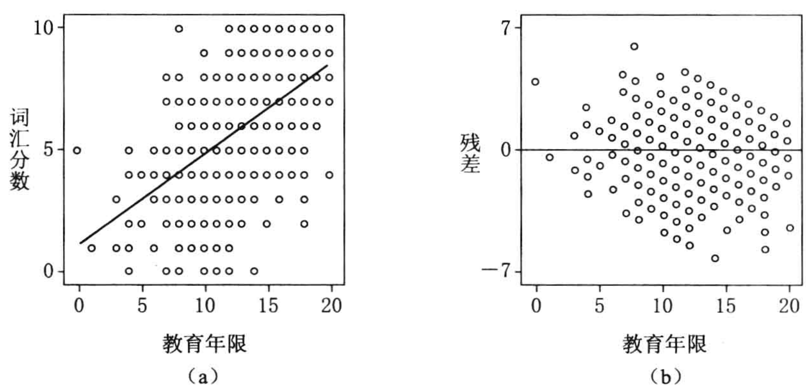
然而在某些情况下，一些转换对于结果的解释具有促进作用。例如，对数转换往往可以获得有意义的解释：$\log_2 x$ 增加 1，则导致 $x$ 翻倍。因此，在重新确认的加拿大职业声望回归中，当使教育和性别构成保持不变时，收入的翻倍将导致职业声望 9 分的增长。

与之类似，面积的平方根或体积的立方根可以被解释为距离或长度的线性测量，跨越一段距离所需时间的倒数则为速度等等。如果 $y$ 与 $x_j$ 都进行了对数转换，则 $x_j'$ 回归系数可以解释为 $y$ 对 $x_j$ 的弹性，也就是说，$x_j$ 1%的改变将对应 $y$ 改变的百分比。在很多情况下，二次方的关系将清晰而有效地解释力度(例如，性别比例适中的职业对声望没有什么影响)，但是四阶的多项式则可能不会。

最后，尽管保持简单性与可解释性很重要，但是没有必要因为坚持用一个明显不适合的函数形式而改变数据。在任何情况下，$y$ 与 $x$ 的拟合关系可用图示或表格表示出来(如果它们经过转换，则使用变量原始的度量)，或者可以描述某一些策略性的 $x$ 值上的效应(例如，上述有关女性比例对职业声望的偏效应的简短描述)。

# 第 8 章 | 离散数据

离散的自变量与因变量往往产生难以解释的散点图。图 8.1 就是这一现象的简单例子，其中的数据来自 1989 年由国家民意研究中心进行的社会概况调查。自变量——教育年限是从 0—20 进行编码的，因变量是在一个 10 个题目的词汇测试中，答对题目的数量。需要注意，这个变量是一个冒充的比例——事实上变量是答对的比例×10。




注：散点图中显示了最小二乘回归直线。

**图 8.1　对词汇分数根据教育年限的散点图与残差散点图**

图 8.1(a)中的散点图与我们的印象一致，随着教育程度的上升，词汇水平也上升。然而这个散点图很难读懂，因为 968 个数据点大多是一个叠加在一个上面的。图中也包括最小二乘回归线，其方程为：

$$\hat{V} = 1.13 + 0.374E$$
$$(0.28)\ (0.021)$$
$$R^2 = 0.248 \quad s = 1.92$$

其中 $V$ 与 $E$ 分别是词汇分数与教育程度。

图 8.1(b)是拟合方程中的残差根据教育程度的散点图。这个图的左下角到右上角的对角线是离散因变量的典型特征：对于 $y$ 的每一个个别值，例如，$y=5$，残差为 $e=5-b_0-b_1x=3.87-0.374x$，是对 $x$ 的线性函数。我注意到，当 $y$ 具有一个既定的最小值时，残差根据拟合值作图可以获得一个与第 6 章类似的现象。从左下角到右上角的对角线是由于 $x$ 的离散性。

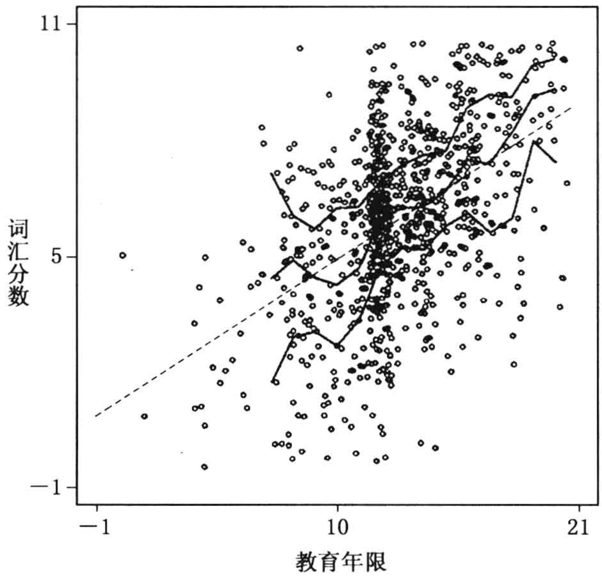
图 8.1(b)显示，对教育这一变量而言，相对于中间值的最大值和最小值都具有较小的残差方差。这一观测的模式与观察到的因变量是一个假比例这一现象一致：当回答正确的平均数量接近 0 或 10 时，词汇分数的潜在变异将降低。然而在这一明显的下降趋势中，部分的原因在于，当接近教育度量的上下限时，数据相对较为分散。我们之所以关注残差值的范围，原因在于我们无法观察到大多数的数据点，而且即便方差是恒定的，这一范围也随着数据量的增加而扩大。

图 8.2 显示了相关的状况，其中每一个数据点都可以水平和垂直地"抖动"，尤其是在每个教育和词汇的分数上加上一个在区间[$-1/2$, $1/2$]的一致随机变量。这种对离散数据绘制散点图的方法是由 Chambers、Cleveland、Kleiner和 Tukey 提出的(1983)。这个散点图同样显示出对原始数据的拟合回归直线，也包括在每个教育的值上抖动的词汇分数的分布上，穿越中位数、第一和第三分位点的直线。我排除了距离中位数与分位点低于 6 的教育值，因为这些数据在这个区域中太过分散。

图 8.2 有很多特点值得关注：(1)从抖动的数据中，我们可以看到观测在教育为 12 年(对应高中毕业)时尤其密集；(2)中位数的轨迹与最小二乘回归直线最为接近；(3)分位点的描记线显示出 $y$ 的扩散程度并没有在教育程度的较大值处降低。

因变量是离散的，违背了回归模型中误差是正态分布且具有一致方差的假设。这个问题与受限因变量一样，只在极端的情况下才会出现。例如，当只有很少的回答类型时，或者很大一部分的观测值包含在很小数目的类别中且取决

于自变量的值之时。




注：对横纵坐标都加上了一个很小的随机量。虚线是对没有抖动数据拟合的最小二乘回归直线。实线是对抖动的词汇分数绘制的中位数和分位数的轨迹。

**图 8.2　对词汇分数根据教育年限的“抖动”散点图**

相比之下，离散的自变量与回归模型完全一致，因为回归模型除了假设 $x$ 与误差不相关之外没有别的对分布的要求。事实上，一个离散的 $x$ 往往可以直接对非线性进行假设检验，往往叫做对“失拟”的检验。同样，对一个离散的自变量的每个类别进行非一致误差方差的检验也相对简单。

## 检验非线性

例如，假设我们将教育转化为一系列虚拟的回归因子进行建模，而非确认词汇分数与教育之间的线性关系。尽管有从 0 到 20 一共有 21 个教育程度的值，但样本中的个体没有一个是具有 2 年教育水平的，这使我们一共有 20 个类别和 19 个虚拟回归因子，则方程变为：

**表 8.1 词汇测试分数不同的分析**

| 出处 | $df$ | 平方和 | 均方 | $F$ | $p$ |
|---|---|---|---|---|---|
| 教育(方程 8.1) | 19 | 1262.0 | 66.40 | 18.1 | $\ll 0.0001$ |
| 线性(方程 8.2) | 1 | 1175.0 | 1175.0 | 320.0 | $\ll 0.0001$ |
| 非性线("拟合不足") | 18 | 86.58 | 4.810 | 1.31 | 0.17 |
| 误差("纯误差") | 948 | 3473.0 | 3.663 | | |
| 总计 | 967 | 4735.0 | | | |

资料来源:1989 年社会概况调查,国家民意调查中心。

$$y_i = \gamma_0 + \gamma_1 d_{1i} + \cdots + \gamma_{19} d_{19,\,i} + \varepsilon_i \qquad [8.1]$$

将这个方程与下面方程相比:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad [8.2]$$

便可以生成一个对非线性的检验,因为方程 8.2 中确认的线性关系是方程 8.1 中的特例,方程 8.1 包含了所有 $E(y)$ 与 $x$ 之间关系的模式。获得的对非线性的增量 $F$ 检验显示在表 8.1 的方差分析中。因此我们清楚,在词汇与教育之间有明显的线性关系,而没有证据显示具有非线性。

假设 $x_1$ 在多元回归模型中,对非线性的 $F$ 检验很容易扩展至离散自变量。这里,我们将较普通的模型,

$$y = \gamma_0 + \gamma_1 d_1 + \cdots + \gamma_{q-1} d_{q-1} + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

与确认了 $x_1$ 线性效应的模型比较:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

其中,$d_1, \cdots, d_{q-1}$ 是构建来表示 $x_1$ 的 $q$ 个类别的虚拟回归因子。

## 检验不一致误差方差

一个离散的 $x$(或几个 $x$ 的组合)将数据分成 $q$ 组。令 $y_{ij}$ 代表在第 $i$ 个组

中 $n_i$ 个因变量分数中的第 $j$ 个。如果误差方差是一致的，则组内的方差估计量

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{n_i - 1}$$

应该是类似的。其中，$\bar{y}_i$ 是第 $i$ 个组的均值。当误差是非正态分布时，直接针对 $s_i^2$ 的检验将不再具有有效性，例如 Bartlett(1937)常用的检验。

因而，许多替代的检验法被提出。在大规模的仿真研究中，Conover、Johnson等人（1981）证明了下面的 $F$ 检验是稳健和有力的：计算 $z_{ij} = |y_{ij} - y_i^*|$ 的值，其中 $y_i^*$ 是第 $i$ 组中 $y$ 的中位数。其后对变量 $z$ 的 $q$ 个组进行单一变量方差分析。如果误差方差在组间是不一致的，则组均值 $\bar{z}_i$ 将有所不同，从而产生一个很大的 $F$ 检验统计值。例如对词汇数据，教育水平将全部 968 个观测分为 $q = 20$ 个组，这个检验得出 $F_{19,948} = 1.48$，$p = 0.08$，并没有产生不一致分布的明显证据。

# 第 9 章 | 最大似然法、计分检验和构造变量

本章介绍的方法都基于最大似然估计法（参见 Fox，1984；Wonnacott & Wonnacott，1990：第 18 章）。这些方法的逻辑比前几章介绍的类似的特定步骤更为复杂，但是应用起来同样很直接。因此，本章的内容对相对没有受到精妙统计理论限制的数据分析者来说，也应该非常有效。

对 $x$ 或 $y$ 转换的选择，一种统计上更复杂的方法是将普通的多元回归模型嵌入一个包含针对转换参数的更一般模型中。如果有好几个变量需要被转换，或者转换非常复杂，则需要有好几个类似的参数。而这一类型的模型本质是非线性的。

假设转换可以由单一参数 $\lambda$ 表示，其后我们记下包含转换参数和普通回归参数的函数形式的模型似然性：$L(\lambda, \beta_0, \beta_1, \cdots, \beta_k, \sigma^2)$。使似然值最大化将获得 $\lambda$ 的最大似然估计（MLE）以及其他参数的最大似然估计。现在，我们令 $\lambda = \lambda_0$，代表没有进行转换（例如 $\lambda_0 = 1$ 表示幂转换 $y^\lambda$）；$H_0$ 的似然比检验：$\lambda = \lambda_0$，用以评估转换是否需要。

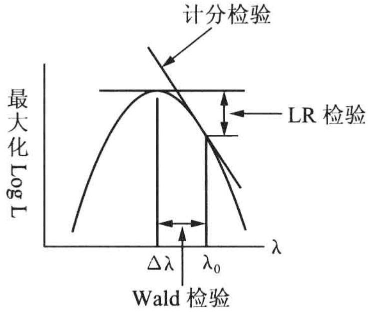



**图 9.1　似然比对假设 $H_0$：$\lambda = \lambda_0$ 的 Wald 和计分检验**

正如图 9.1 中所示，似然比检验比较 MLE$\hat{\lambda}$ 的对数似然值和零假设值 $\lambda_0$：如果 $\log_e L(\hat{\lambda})$ 比 $\log_e L(\lambda_0)$ 大很多，$H_0$ 将被拒绝，则我们可以获得结论——需要进行转换。如图 9.1 中所示的替代检验就是基于 $\hat{\lambda}$ 与 $\lambda_0$ 之间距离的 Wald 检验；计分检验（也称做"朗格朗日乘数递增检验"）则基于 $\lambda_0$ 处似然对数的斜率——一个对 $H_0$ 产生质疑的

陡峭斜率，因为在最大处（即当 $\lambda = \hat{\lambda}$ 时）的对数似然是平坦的。对于二次项的对数似然值，这 3 个检验是同样的，但是在更一般的时候不是，尽管在实践中，它们往往产生类似的 $p$ 值并具有渐近（大样本）的特质。

似然比率与 Wald 检验的优点在于它们需要找到 MLE，而这往往需要迭代（一个不断近似的重复过程）。相比之下，在 $\lambda_0$ 处，$\log_e L$ 的斜率往往不需要迭代就可以被估计或者近似。计分检验可以构造成对一个新回归因子的 $t$ 统计量，称做“构造变量”，然后将其加入回归模型。此外，对构造变量的偏回归散点图能够揭示一个或者一部分观测是否会产生严重的影响、是否进行转换，或者是否要对整个数据进行转换。

## y 的 Box-Cox 转换

Box 与 Cox 建议对 $y$ 进行幂转换以使误差呈正态分布，使误差方差变稳定并使 $y$ 与 $x$ 之间的关系变为线性。一般的模型为：

$$y_i^{(\lambda)} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

$$\varepsilon_i \sim \mathrm{NID}(0,\ \sigma^2)$$

其中

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} (\lambda \neq 0) \\ \log_e y_i (\lambda = 0) \end{cases}$$

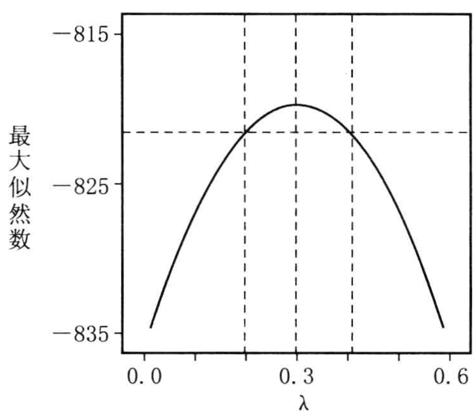
其中，所有的 $y_i$ 都是正的。对某一选定的 $\lambda$，Box 与 Cox 证明条件的最大化对数似然值为：

$$\log_e L(\beta_0,\ \beta_1,\ \cdots,\ \beta_k,\ \sigma^2 \mid \lambda) = -\frac{n}{2}(1 + \log_e 2\pi) - \frac{n}{2}\log_e s^2(\lambda) + (\lambda - 1)\sum_{i=1}^{n} \log_e y_i$$

其中，$s^2(\lambda)=\sum e_{(\lambda)i}^2/n$，且 $e_{(\lambda)i}$是对 $y^{(\lambda)}$ 根据 $x$ 进行最小二乘线性回归得到的残差。寻找 MLE$\hat{\lambda}$ 的一个简单步骤就是评估某一段 λ 值(例如在 −2 到 +2 之间)的最大 $\log_e L$。如果结果显示这段区间没有包括似然对数的最大值，则需要扩大这段区间。检验 $H_0:\lambda=1$，则需要计算似然比检验统计量：

$$G_0^2=-2\times[\log_e L(\lambda=1)-\log_e L(\lambda=\hat{\lambda})]$$

在 $H_0$ 情况下是 $\chi_1^2$ 分布。对 λ 的 95%置信区间包括那些符合 $\log_e L(\lambda)>\log_e L(\lambda=\hat{\lambda})-1/2\times1.96^2$ 的值，其中 $1.96^2=\chi_{1,0.05}^2$。图 9.2 显示的是针对 Ornstein 连锁企业董事会回归所做的、针对 λ 的最大似然对数的散点图。λ 的最大似然估计是 $\hat{\lambda}=0.30$，而 95%的置信区间则从 0.20 到 0.41，图中在接近似然对数处用交叉线标示出来(在第 6 章中，我们对这一数据使用了平方根转换以使误差方差稳定化)。




注：在 Box-Cox 模型中作为对参数 λ 转换函数的条件最大似然对数。图上方水平线与似然对数的交叉组成了 λ 的 95%置信区间。

**图 9.2 对 Ornstein 的连锁董事会回归**

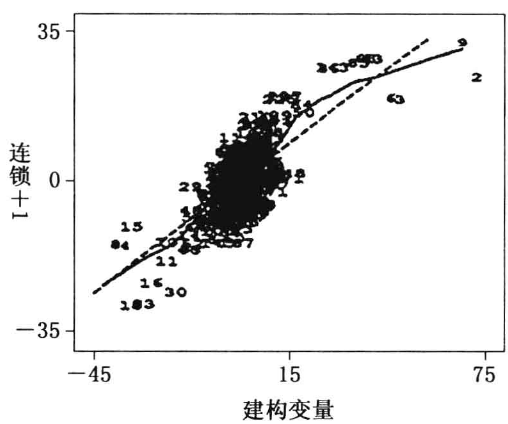
Atkinson(1985)针对 Box-Cox 模型提出了一个近似的计分检验，这一检验基于构造变量 $G_i=y_i\times[\log_e(y_i/\tilde{y})-1]$，其中 $\tilde{y}$ 是 $y$ 的几何平均数，$\tilde{y}=(y_1\times$

$y_2 \times \cdots \times y_n)^{1/n}$。这一构造变量是通过对在 $\lambda = 1$ 处的 Box-Cox 转换 $y^{(\lambda)}$ 的线性近似获得的。扩展方程为：

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varphi G_i + \varepsilon_i$$

对 $H_0: \varphi = 0$ 的 $t$ 检验即 $t_0 = \hat{\varphi}/SE(\hat{\varphi})$，并通过此来评估是否需要进行转换。对 $\lambda$ 的估计（尽管不是 MLE）为 $\tilde{\lambda} = 1 - \hat{\varphi}$，而对 $G$ 的偏回归散点图则表示了对 $\hat{\varphi}$ 的影响，以及对 $\lambda$ 的选择。

图 9.3 是对连锁企业董事会回归的 Atkinson 构造变量散点图。尽管图中的趋势并非始终为线性，但可以看出，对 $y$ 的转换是需要在整个数据中进行的，且并非源于一小部分影响力很大的观测。构造变量的回归系数为 $\hat{\varphi} = 0.588$ 且 $SE(\hat{\varphi}) = 0.032$，强烈表明需要对 $y$ 进行转换。而建议的转换 $\tilde{\lambda} = 1 - 0.588 = 0.412$ 与 MLE 非常接近。




注：图中每个点都标有观测索引。最小二乘（虚线）和 lowess 回归（实线）均显示在图中。

**图 9.3　对 Ornstein 连锁董事会回归的 Box-Cox 转换的建构变量散点图**

## 对 $x$ 的 Box-Tidwell 转换

现在我们考虑如下模型：

$$y_i = \beta_0 + \beta_1 x_{1i}^{\gamma_1} + \cdots + \beta_k x_{ki}^{\gamma_i} + \varepsilon_i$$

$$\varepsilon_i \sim \text{NID}(0,\ \sigma^2)$$

在假设模型中，所有 $x_{ji}$ 都是正的，则这个模型中的参数，$\beta_0$，$\beta_1$，…，$\beta_k$ 和 $\gamma_1$，…，$\gamma_k$ 可以与 $\sigma^2$ 一起，通过一般非线性最小二乘模型(参见 Gallant，1975)估计出来，但是 Box 与 Tidwell(1962)提出一种更有效的步骤来进行构造变量的诊断：

第一，将 $y$ 根据 $x_1$，…，$x_k$ 进行回归，得到 $b_1$，…，$b_k$。

第二，将 $y$ 根据 $x_1$，…，$x_k$ 与构造变量 $x_1\log_e x_1$，…，$x_k\log_e x_k$ 进行回归，得到 $b'_0$，$b'_1$，…，$b'_k$ 与 $d_1$，…，$d_k$。注意，由于在第二个回归中加入了构造变量，一般来讲 $b_j \neq b'_j$。正如在 Box-Cox 模型中，构造变量是在 $\gamma=1$ 处对 $x_j^\gamma$ 的线性近似得到的。

第三，根据对假设 $H_0$：$\delta_j=0$ 的检验，构造变量 $x_j\log_e x_j$ 可用于评估是否需要对 $x_j$ 进行转换，其中 $\delta_j$ 是第二个回归中 $x_j\log_e x_j$ 的整体系数。构造变量的偏回归散点图对于评价对 $x$ 进行转换的影响力与影响程度非常有效。

第四，对 $\gamma_j$ 的估计可以通过 $\tilde{\gamma}_j = 1 + d_j/b_j$ 获得。$b_j$ 可以通过第一步回归获得。

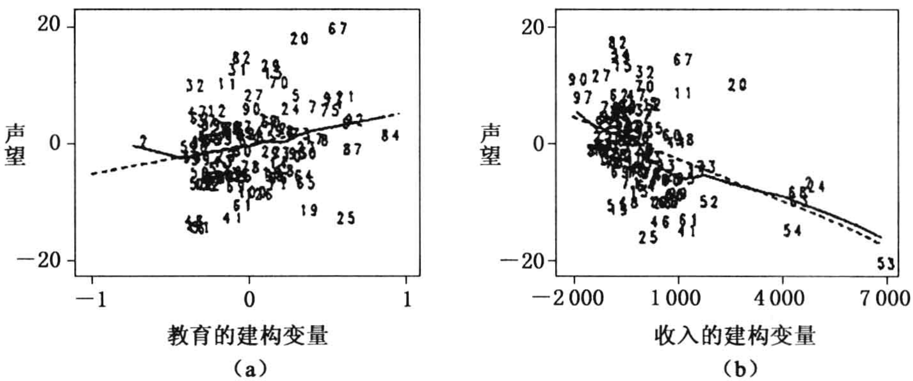
可以通过重复进行第一、第二和第四步来完成这一程序，直到对转换参数的估计值稳定下来，获得 MLE 的 $\hat{\gamma}_j$。

对于加拿大的职业声望数据，保持女性比例这一变量不变($W$ 与 $W^2$)，则在辅助回归中，$E\log_e E$ 与 $I\log_e I$ 的系数分别为 $d_E = 5.30$ 且 $\text{SE}(d_E) = 2.20$，而 $d_I = -0.00243$ 且 $\text{SE}(d_I) = 0.00046$。这一结果表明，相对于教育，我们更需要对收入进行转换。回顾第 7 章我们发现，对教育的幂转换并非十分合适。转换参数的第一步估计为：

$$\tilde{\gamma}_E = 1 + d_E/b_E = 1 + 5.30/4.26 = 2.2$$

$$\tilde{\gamma}_I = 1 + d_I/b_I = 1 - 0.00243/0.00127 = -0.91$$

对转换参数进行完全的 MLE 迭代，得到 $\hat{\gamma}_E = 2.2$ 且 $\hat{\gamma}_I = -0.038$。将这一结果与第 7 章中通过反复试错法得出的平方与对数转换相比较。从图 9.4 中对转换的教育与收入的构造变量散点图可以看出转换大体上是必要的，除了在收入的散点图上有一些高影响力的观测值。




注：每个点都标有观测索引。最小二乘（虚线）和 lowess 回归（实线）均显示在图中。

**图 9.4 在加拿大职业声望回归中，对教育和收入进行 Box-Tidwell 转换后的建构变量散点图**

## 对不一致误差方差的矫正

对于异方差问题，Breusch 与 Pagan(1979)提出了一种计分检验，它基于确认误差方差与已知变量 $z_1, \cdots, z_p$ 相关而获得，并且可以建模为：

$$\sigma_i^2 = V(\varepsilon_i) = g(\gamma_0 + \gamma_1 z_i + \cdots + \gamma_p z_{pi})$$

其中，函数 $g(\cdot)$非常普通，而类似的检验也由 Cook 与 Weisberg(1983)独立提出。对于 $\sigma^2$ 恒定的这一假设（与 $H_0: \gamma_1 = \cdots = \gamma_p = 0$ 等同）的计分检验，可以转换成一个辅助回归的问题。

令 $u_i = e_i^2/\hat{\sigma}^2$，其中 $\hat{\sigma}^2 = \sum e_i^2/n$ 是对误差方差的 MLE（注意除数为 $n$ 而不是自由度 $n-k-1$），$u_i$ 为一种标准化残差的平方。将 $u$ 根据 $z$ 进行回归：

$$u_i = \alpha_0 + \alpha_1 z_{1i} + \cdots + \alpha_p z_{pi} + \omega_i \quad [9.1]$$

Breusch 和 Pagan 证明了计分统计量 $S^2 = \sum(\hat{u}_i - \bar{u})^2/2$ 在 $H_0: \sigma_i^2 = \sigma^2$ 时近似服从 $\chi_p^2$ 分布。这里，$\hat{u}_i$ 是将 $u$ 根据 $z$ 进行回归而获得的拟合值，而 $S^2$ 则为根据拟合方程 9.1 获得的回归平方和的一半。

在应用中，当然需要选择 $z$，而这种选择则基于对不一致误差方差模式的预估。如果预估了好几种模式，则需要进行多个计分检验。例如在辅助回归方程 9.1 中使用 $x_1$，…，$x_k$，则允许在主要回归中发现误差方差随着一个或一个以上自变量的增加而增加这一趋势。

与之类似，Cook 与 Weisberg(1983)提出了将 $u$ 根据从主要回归中获得的拟合值进行回归，从而得到一个自由度为 1 的计分检验，用以探测误差方差随着因变量增大而增加的一般趋势。当事实上误差方差正符合这一模式时，与更一般的对 $u$ 根据 $x$ 进行回归的结果相比，对 $u$ 根据 $\hat{y}$ 进行辅助回归将获得更有效的检验。Anscombe 提出了一个类似(但更复杂)的步骤，他建议利用将 $y$ 转换为 $y^{(\tilde{\lambda})}$ 来矫正已发现的异方差，其中 $\tilde{\lambda} = 1 - 1/2\hat{\alpha}_1\bar{y}$。

最后，White(1980)提出了一个类似的计分检验法，这种方法基于将他自己提出的系数抽样方差的异方差矫正估计量(参见第 6 章与附录 9)与一般的系数方差估计量进行的比较。如果两个估计量差异很大，则需要对一致误差方差这一假设提出怀疑。White 的检验可以作为对从主要回归中获得的残差平方进行的辅助回归，即对 $e^2$ 根据所有的 $x$ 及其平方和它们之间的两两组合。这样，对于包含两个自变量的主要回归，我们可以拟合模型为：

$$e_i^2 = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \delta_3 x_{1i}^2 + \delta_4 x_{2i}^2 + \delta_5 x_{1i} x_{2i} + v_i$$

一般来说，除了常数项之外，辅助回归中还有 $p = k(k+3)/2$ 个项。

检验零假设(一致的误差方差)的计分统计量为 $S^2 = nR^2$，其中 $R^2$ 是辅助回归中复相关系数的平方。在零假设成立的情况下，$S^2$ 服从自由度为 $p$ 的近似 $\chi^2$ 分布。

由于所有计分检验都很容易违背除了一致误差方差之外的回归的其他假设，因此在实践中应该利用图示来辅助这些检验(Cook & Weisberg, 1983)。当存在多个 $z$ 时，一个简单的诊断则为对 $u_i$ 根据 $\hat{u}_i$ 绘制散点图，$\hat{u}_i$ 是辅助回归中

的拟合值。当将 $\hat{u}_i$ 简单地根据 $\hat{y}_i$ 进行回归，则获得的散点图与第 6 章中提到的根据拟合值对学生残差进行回归获得的散点图非常相似。

根据 Ornstein 的连锁企业董事会数据，对 $u$ 根据 $\hat{y}$ 进行的辅助回归将得到 $\hat{u}=0.134+0.0594\hat{y}$，以及一个自由度下的 $S^2=147.6/2=73.8$。因此，有非常有力的证据显示误差方差是不一致的。利用 Anscombe 的方法进行误差稳定化的转换为 $\tilde{\lambda}=1-1/2(0.0594)(14.81)=0.56$。将这个值与 Box-Cox 模型（$\hat{\lambda}=0.3$）、反复试错法（$\lambda=0.5$，见第 6 章）获得的值进行比较。

对 $u$ 根据主要回归中的自变量进行的辅助回归，可以得到在自由度为 $k=13$ 时，$S^2=172.6/2=86.3$，这也同样为误差方差不一致提供了有力的证据。检验辅助回归的系数可以发现，随着公司资产的增大，误差方差随之将增加这一趋势。然而需要注意的是，对于更一般的检验，计分统计量与对 $u$ 根据 $\hat{y}$ 进行回归获得的结果相差不大，这表明，不一致误差方差的模式的确是方差的分布随着 $y$ 的增大而扩散。公司资产当然是 $\hat{y}$ 的一个重要组成部分。因为 White 的检验需要 104 个回归因子，因此在此并不显示。

# 第 10 章 | 建 议

1. 在进行复杂的统计分析之前，先过滤你的数据。检验单一变量分布和双变量散点图尽管并不能取代本书中提供的方法，但可以揭示出类似奇异数据值、高度偏斜的分布、极端的非线性等等。如果数据集很小，则可考虑自己将数据输入电脑。一般来说，着手处理数据不要犹豫。

2. 当然需要利用一小部分简单稳健、信息量较足的诊断方法，而对需要用更复杂方法才能揭示出的问题，必须追查到底。下面对常见诊断方法的选择建议非常有用：

(1) 共线性：尽管共线性对于个体层面的截断数据并不是一个严重的问题(对于汇总或追踪数据则更常见)，但可以简单地通过计算方差膨胀因子进行诊断。

(2) 强影响数据、奇异值与非正态：除非是总体不准确的数据(例如，将缺失值编码为有效数据)，与较大的数据集相比，强影响数据往往出现在较小的数据集中。一个对学生残差根据预测值绘制的散点图就是一个很好的诊断，因为我提及的所有的影响程度统计量都或多或少取决于相关的值。Cook 的 $D$ 的索引散点图为回归系数的影响程度提供了一个概括性的测量。偏回归散点图对于显示对单一系数的影响力与影响程度非常有效，并且可能揭示出一些强影响的观测子集，而这在单一观测删除统计量中则无法实现。学生残差的正态分位数比较散点图则可以揭示出奇异值以及偏斜与重尾分布。茎叶图、直方图或者学生残差的平滑直方图则可以显示出残差分布的形状，并可能揭示出多路方式这样的问题。

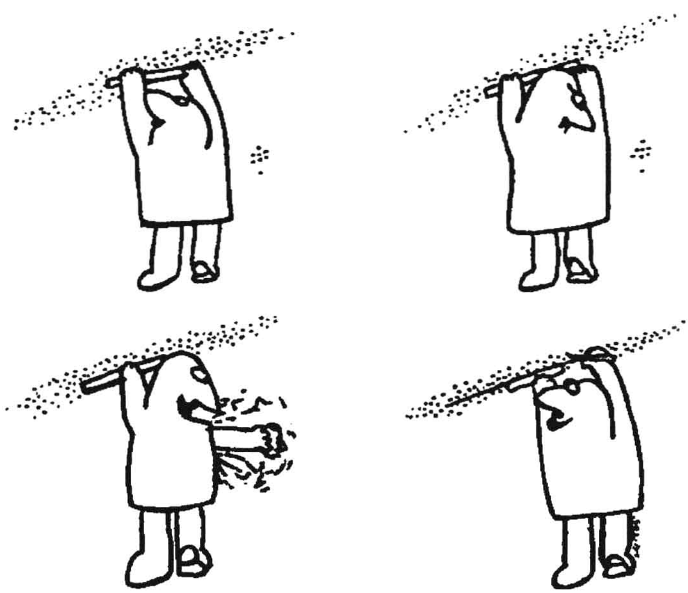
(3) 非线性：如果在偏回归散点图中显示出了非线性，那么在偏残差散点图中这一趋势将更加明显。因为后者非常容易构建，所以可以经常使用。

(4) 不一致误差方差：对学生残差根据拟合值绘制的散点图将会揭示误差方差随着 $y$ 的等级而改变这一趋势，而这也是这一类型问题中的典型。

3. 如果可以的话，尽量使用平滑的方法（例如 lowess 平滑散点图）以揭示出数据的模式。尽管诊断的技术往往只是指示大体的方向而不是细节的问题，但有时也需要对此进行弥补，使我们能够感知到一些看不见的视觉模式并分离出一些视觉的干扰。

4. 尽量避免“过度拟合”这一陷阱（例如，对统计模型进行修改以捕捉数据中的偶然特征）。数据分析的部分艺术就在于判断如何根据数据进行调试。一个极端情况是，一些人忽略了数据中无法预知的模式来对教科书进行模仿以获得“客观”的统计分析，而这就要求模型是事先被确定好的。而另一极端则是，那些诊断技术的初学者往往删除了大部分数据，或者不停地对数据进行转换以获得微小的“较好”拟合。

图 10.1 挖苦了对不一致数据的抛弃。我在此必须指出，对所有的数据勉强拟合出一条直线，比删除那些令人恼怒的数据点更糟糕。当然，最好是能够理解这部分不一致的数据与其余数据有什么不同。



资料来源：获得大学联合会政治和社会研究暑期课程授予的重新印刷权利，1990 年。

**图 10.1 行动中的回归诊断**

尽管对“确认搜索”(即通过检验数据而选择统计模型)的正规统计分析是非常复杂的(参见 Leamer, 1978),但是通过交互效度来评估模型的完备性是可行的(Mosteller & Tukey, 1977)。为了使结果具有交互效度,我们首先将样本随机分为两个部分(并不一定要求分成同等规模)。一个子样本用来根据数据选择一个模型,然后利用另一半数据来测量这一模型的效度。当根据数据选择回归中的一部分自变量时,或当使用转换以应对非线性问题时,这一方法尤为有效。交互效度对于奇异与强影响数据并没有直接的作用,这些问题都是基于个体数据,而不是两个子数据而产生的。

交互效度要求的将样本分开,但研究者往往不愿意这样做,因为他们对于样本规模对估计精度的影响以及统计检验力度的影响非常敏感。但是,当数据的一部分用来选择模型,另一部分用于估计参数时,估计的精度是不切实际的,至少一部分是这样。然而,在我看来更糟糕的是,避免检验一个最先确认的模型的完备性,只是简单地为了保护经典的估计和检验“不受污染”。

正如上文提到的,删除奇异和强影响值并不能赋予数据交互效度。但是,在此之后对抽样方差的估计则是趋近最优的。然而,如果用于拒绝奇异和强影响数据的法则可以被精确地说明,那么可以在这之后对抽样方差进行估计(参见 Diaconis & Efron, 1983; Stein, 1965)。

5. 考虑数据的抽样特征。基于复杂的抽样设计获得的数据往往在观测中有不可忽略的属性(例如 Kish, 1965)。同样,大量的缺失数据则需要特殊的处理办法(例如 Little & Rubin, 1990)。

误差独立这一假设往往不切实际,这一情况常常发生在当观测是由时间点进行划分之时,这也产生了我们称之为“时间序列”的数据。用于探测和解决时间序列回归中误差自相关问题的方法,可以在 Ornstein(1990)和 Kmenta(1986)等书中找到。在这种情况下,一个有用的初步诊断就是对最小二乘残差根据代表时间的观测索引绘制散点图。

## 计算诊断量

现在标准的统计软件(例如 SAS、SPSS、BMDP 和 Systata)已经包含了许

多本书讨论的诊断方法。哪怕某一个特殊的统计量或者步骤不是直接由这一软件提供,但常常也很容易计算或构造。例如在第 4 章中讨论到 DFFITS 和 COVRATIO,可以根据预测值和学生残差获得。同样,第 4 章与第 9 章中讨论的偏回归与构造变量散点图,可以通过构造某个适合的回归的残差而获得。即便是相对复杂的步骤,例如 lowess,也可以通过在 SAS 里进行编程而获得(而事实上,lowess 的散点图平滑方法在 Systat 和其他软件中已经包括)。

本书中几乎所有的计算都是使用 SAS 的 PC 版获得的,图形(仅有少数不是)则是由 SAS/GRAPH 获得。目前的统计软件对于应用诊断技术并不困难,并且标准软件的诊断功能也将不断进步。

## 延伸阅读

对于回归诊断和相关的主题,有大量的文献材料,例如"探索与图示数据分析"。幸运的是,现在有许多相关的文献采用了更易阅读的形式。

在我看来,Cook 和 Weisberg(1982b)的书对于了解评估影响力(作者并不喜欢这个词语)、奇异值和影响程度是最好的。这本书同样包括了对其他问题的讨论,例如非线性与对自变量和因变量的转换,但是并不包括对于共线性的处理。Cook 和 Weisberg(1982a)的文章则以较浓缩的模式讨论了本书中出现的主题。

Chatterjee 和 Hadi(1988)的书是一本关于处理强影响数据的全面且新近的著作,此外,这本书也讨论了非线性与非一致误差方差等问题。这本书的杰出之处在于对不同测量影响程度的回归结果进行了比较,包括回归系数、系数方差以及共线性。

Atkinson(1985)的书同样是一本非常有价值的著作,这本书强调了作者对回归诊断的重要贡献,例如构造变量散点图和仿真的方法。Belsley 等人(1980)的著作处理了强影响数据与共线性,主要介绍了作者在这些领域的工作。然而我认为,他们对共线性的处理都因主张在评估有问题的情况之前,不应该将截距处理掉这一主张而有所缺憾(参见 Belsley, 1984)。

一些应用回归和线性模型的书对于诊断有较多的处理。例如 Chatterjee 和 Price(1977)、Daniel 和 Wood(1980)以及 Draper 和 Smith(1981)对于共线性、变量选择和其他基于残差的诊断方法的讨论。此外还有 Weisberg(1985)、Fox(1984)对于本书中的一些主题的处理。一般讨论计量的著作包含探测和矫正对回归模型假设的违背,但往往是以理论为主而非基于数据分析。对于这一方法的例子,参见 Kmenta(1986)。

对于数据分析的图示和探测法,有许多优秀的著作,包括 Cleveland(1985)对绘图的介绍,Velleman 和 Hoaglin(1981)对探索数据分析的介绍。同样参见 Chambers 等人(1983)的著作,其中包括与 Cleveland 类似的内容;Tukey(1977)的著作包括对于探索数据的分析;由 Hoaglin、Mosteller 和 Tukey 编辑的丛书(1983、1985)以及 Fox 和 Long 编辑的丛书(1990)也提供了有用的帮助。最后,Mosteller 和 Tukey(1977)作为对 Tukey(1977)手册的特殊回归著作,从数据分析的视角研究了许多与回归相关的有趣内容。

# 附 录

## 附录 1　最小二乘拟合、联合置信区域和检验

利用矩阵形式，线性回归模型可以被写成 $\mathbf{y}=\mathbf{X}\beta+\varepsilon$，其中 $\mathbf{y}$ 是一个 $n\times 1$ 维的由因变量值组成的向量。$\mathbf{X}$ 是 $n\times(k+1)$ 的回归因子矩阵，其中包括常量回归因子，即矩阵中全部为 1 的第一列；$\beta$ 是 $(k+1)\times 1$ 的回归参数向量，$\varepsilon$ 是 $n\times 1$ 的误差向量。根据回归假设，$\varepsilon\sim N_n(0,\ \boldsymbol{\sigma}^2\mathbf{I})$，且与 $\mathbf{X}$ 独立。

拟合模型为 $\mathbf{y}=\mathbf{X}\mathbf{b}+\mathbf{e}$。为了获得 $\beta$ 的最小二乘估计 $\mathbf{b}$，我们需要使得残差的平方和最小，$\mathbf{e}'\mathbf{e}=[\text{length}(\mathbf{e})]^2$。因为 $\mathbf{e}=\mathbf{y}-\hat{\mathbf{y}}$，通过使 $\hat{\mathbf{y}}=\mathbf{X}\mathbf{b}$ 成为 $\mathbf{y}$ 在由 $\mathbf{X}$ 的列获得的子空间上的垂直投影，使 $\mathbf{e}$ 的长度最小化。由于 $\mathbf{X}'\mathbf{e}=\mathbf{0}$，我们有 $\mathbf{X}'\mathbf{X}\mathbf{b}=\mathbf{X}'\mathbf{y}$，这是矩阵形式的一般方程。需要注意的是，由于 $\hat{\mathbf{y}}$ 在 $\mathbf{X}$ 的子空间的列上，所以残差与拟合值是垂直的：$\sum e_i\hat{y}_i=\mathbf{e}'\hat{\mathbf{y}}=0$。此外，由于 $\mathbf{X}$ 的第一列为 1，则有 $\sum e_i=\mathbf{1}'\mathbf{e}=0$。

另一种等价的表达为：

$$\mathbf{e}'\mathbf{e}=(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})=\mathbf{y}'\mathbf{y}-2\mathbf{y}'\mathbf{X}\mathbf{b}+\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

求微分可获得 $\partial\mathbf{e}'\mathbf{e}/\partial\mathbf{b}=-2\mathbf{X}'\mathbf{y}+2\mathbf{X}'\mathbf{X}\mathbf{b}$，通过求偏导可以使推导出一般方程的平方和函数最小化。如果 $\mathbf{X}'\mathbf{X}$ 是非奇异的，即 $\mathbf{X}$ 的列中不存在共线性，则有 $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$。

根据假设，我们有 $E(\varepsilon)=\mathbf{0}$，且 $E(\mathbf{y})=\mathbf{X}\beta$，$E(\mathbf{b})=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y})$，则 $\mathbf{b}$ 为 $\beta$ 的无偏估计。根据假设，我们有 $V(\mathbf{y})=V(\varepsilon)=\sigma^2\mathbf{I}$，利用平方与乘积之和的矩阵 $\mathbf{X}'\mathbf{X}$ 的对称性，

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

根据误差正态分布这一假设，则有：

$$\mathbf{b} \sim N_{k+1}[\beta,\ \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$$

则对回归系数的 $100(1-\alpha)\%$ 的椭圆联合置信区域为：

$$(\mathbf{b}-\beta)'(\mathbf{X}'\mathbf{X})(\mathbf{b}-\beta) \leqslant (k+1)s^2F_{\alpha,\ k+1,\ n-k-1}$$

其中，$s^2 = \mathbf{e}'\mathbf{e}/(n-k-1)$ 是对 $\sigma^2$，$F_{\alpha,\ k+1,\ n-k-1}$ 对于拥有 $k+1$ 和 $n-k-1$ 自由度的 $F$ 的临界值。对于 $p$ 个回归参数中的 $\beta_1$，我们有 $100(1-\alpha)\%$ 的置信区域：

$$(\mathbf{b}_1-\beta_1)\mathbf{V}_{11}^{-1}(\mathbf{b}_1-\beta_1) \leqslant ps^2F_{\alpha,\ p,\ n-k-1} \qquad [\text{A. }1]$$

在这里，$\mathrm{V}_{11}$ 是 $(\mathbf{X}'\mathbf{X})^{-1}$ 相对于 $\mathbf{b}_1$ 的行与列的 $p\times p$ 子阵。

通过置信区域的表达式，可以很容易地进行 $F$ 检验。例如，检验 $H_0:\beta_1=\beta_1^{(0)}$，则有：

$$F_0 = \frac{(\mathbf{b}_1-\beta_1^{(0)})'\ \mathbf{V}_{11}^{-1}(\mathbf{b}_1-\beta_1^{(0)})}{ps^2}$$

在 $H_0$ 假设下服从 $F_{p,\ n-k-1}$ 分布。对于 $\beta_1^{(0)}=\mathbf{0}$，$F_0$ 即为增量 $F$ 统计量。

## 附录 2 岭回归

岭回归(Hoerl & Kennard，1970a、1970b)是在具有较强共线性的情况下获得更有效估计的一种方法。在这里，我解释岭回归的首要目的就是提醒大家，岭回归并不是对共线性的一种一般补救方法。

通过重新度量 $\mathbf{y}$ 和 $\mathbf{X}$ 的列，使它们的均值为 0，且具有单位长度，则求和即可得到相关系数。对于标准化回归系数的岭估计为：

$$\mathbf{b}_z^* = (\mathbf{R}_{xx}z\mathbf{I})^{-1}\mathbf{r}_{xy} = (\mathbf{R}_{xx}+z\mathbf{I})^{-1}\mathbf{R}_{xx}\mathbf{b}^*$$

其中，$\mathbf{b}^* = \mathbf{R}_{xx}^{-1}\mathbf{r}_{xy}$ 是最小二乘估计量，$z \geqslant 0$ 是岭常量，通常由研究者自己选

定。在这里，$\mathbf{R}_{xx}$ 是 $x$ 之间的相关系数矩阵，$\mathbf{r}_{xx}$ 是 $x$ 与 $y$ 之间相关系数的矢量。通过对每个 $\mathbf{R}_{xx}$ 的对角值加上 $z$，则对角值(原来为 1)比非对角值(回归因子间的相关系数)膨胀了一部分，从而提高了自变量相关系数矩阵的调节。当 $z = 0$ 时，最小二乘和岭估计量相等：$\mathbf{b}_0^* = \mathbf{b}^*$。

Hoerl 和 Kennard 证明了 $\mathbf{b}_z^*$ 的偏差随着 $z$ 的增加而增加，即在 $z > 0$ 时，$V(\mathbf{b}_z^*) < V(\mathbf{b}^*)$，且 $V(\mathbf{b}_z^*)$ 随着 $z$ 的增大而减小，那么 $z$ 永远存在一系列值使 $\mathrm{MSE}(\mathbf{b}_z^*) < \mathrm{MSE}(\mathbf{b}^*)$。之前提到的均方误即抽样方差和偏差平方之和，则岭回归中的技巧就是选择使偏差与方差之间权衡最优的 $z$ 值。

评断 $z$ 值为多少可以使岭估计量优于最小二乘估计量，这取决于未知参数 $\beta^*$，因此在实践中无法看出岭估计量具有何种理论上的优势。

## 附录 3　预测值和预测矩阵

最小二乘回归的拟合值是观测 y 的线性函数：

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

这里，$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 就是预测矩阵。这样命名的原因在于它将 $\mathbf{y}$ 转换为 $\hat{\mathbf{y}}$。预测矩阵是对称($\mathbf{H} = \mathbf{H}'$)和幂等($\mathbf{H}^2 = \mathbf{H}$)的，这点很容易被证明。因此预测矩阵的对角值 $h_i = h_{ii}$ 叫做“预测值”，即

$$h_i = \sum_{j=1}^{n} h_{ij}^2 = h_i^2 + \sum_{j \neq i} h_{ij}^2$$

且有 $0 \leqslant h_{ii} \leqslant 1$。如果 $\mathbf{X}$ 包括了常数回归因子，则有 $1/n \leqslant h_i$。最后，由于 $\mathbf{H}$ 是一个投影矩阵，将 $\mathbf{y}$ 正交地投影在有 $\mathbf{X}$ 的列组成的子空间上，则有 $\sum h_i = k+1$，因此 $\bar{h} = (k+1)/n$。细节参见 Hoaglin 和 Welsch(1978)或者 Chatterjee 和 Hadi(1988:第 2 章)。

## 附录 4　最小二乘残差的分布

最小二乘的残差为：

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{X}\beta + \varepsilon) - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) = (\mathbf{I} - \mathbf{H})\varepsilon$$

因此有

$$E(\mathbf{e}) = (\mathbf{I} - \mathbf{H})E(\varepsilon) = (\mathbf{I} - \mathbf{H})\mathbf{0} = \mathbf{0}$$

与

$$V(\mathbf{e}) = (\mathbf{I} - \mathbf{H})V(\mathbf{e})(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})$$

因为 $\mathbf{I} - \mathbf{H}$ 与 $\mathbf{H}$ 一样是对称和幂等的。$\mathbf{I} - \mathbf{H}$ 不是对角矩阵，且其对角值往往是不相等的，因此即便误差服从独立同方差假设，残差之间仍彼此相关且具有不同的方差。

## 附录 5　删除诊断量

令 $\mathbf{b}_{(-i)}$ 表示忽略了第 $i$ 个观测的最小二乘回归系数的矩阵。则 $\mathbf{d}_i = \mathbf{b} - \mathbf{b}_{(-i)}$ 表示了第 $i$ 个观测对回归系数的影响，$\mathbf{d}_i$ 可以由以下方程算出：

$$\mathbf{d}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{e_i}{1 - h_i} \qquad [A.2]$$

Cook 的 $D_i$ 是对“假设”$\beta = \mathbf{b}_{(-i)}$ 检验的 $F$ 值：

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{(-i)})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}_{(-i)})}{(k+1)S^2} = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})}{(k+1)S^2}$$

因此，另一个对 $D_i$ 的解释就是它测量了观测 $i$ 对拟合值 $\hat{\mathbf{y}}$ 的汇总影响，这就是 Belsley 等人(1980)称他们的类似统计量为“DFFITS”的原因。利用方程 A. 2 可得：

$$D_i = \frac{e_i^2}{s^2(k+1)} \times \frac{h_i}{(1 - h_i)^2} = \frac{e'^2_i}{k+1} \times \frac{h_i}{1 - h_i}$$

这也是书中所给出的方程。

## 附录 6　偏回归散点图

用矩阵的形式，拟合的多元回归模型为：

$$\mathbf{y} = b_0\mathbf{1} + b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k + \mathbf{e} \qquad [\text{A. }3]$$

其中，$\mathbf{y}$ 和 $\mathbf{x}_j$ 是观测的 $n\times 1$ 维向量，$\mathbf{1}$ 是 $n\times 1$ 维向量。在最小二乘回归中，是 $\hat{\mathbf{y}} = b_0\mathbf{1} + b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k$ 是 $\mathbf{y}$ 在回归子空间的正交投影。令 $\mathbf{y}^{(1)}$ 和 $\mathbf{x}^{(1)}$ 分别表示 $\mathbf{y}$ 与 $\mathbf{x}_1$ 在由 $\mathbf{1}$ 和 $\mathbf{x}_2$，…，$\mathbf{x}_k$ 组成的子集的补集上的投影(即对 $y$ 和 $x_1$ 根据其他 $x$ 回归获得的残差向量)。根据投影的几何性质，$\mathbf{y}^{(1)}$ 在 $\mathbf{x}^{(1)}$ 上的投影为 $b_1\mathbf{x}^{(1)}$，且 $\mathbf{y}^{(1)} - b_1\mathbf{x}^{(1)} = \mathbf{e}$，即从方程 A. 3 中获得残差向量。

## 附录 7　lowess 平滑散点图

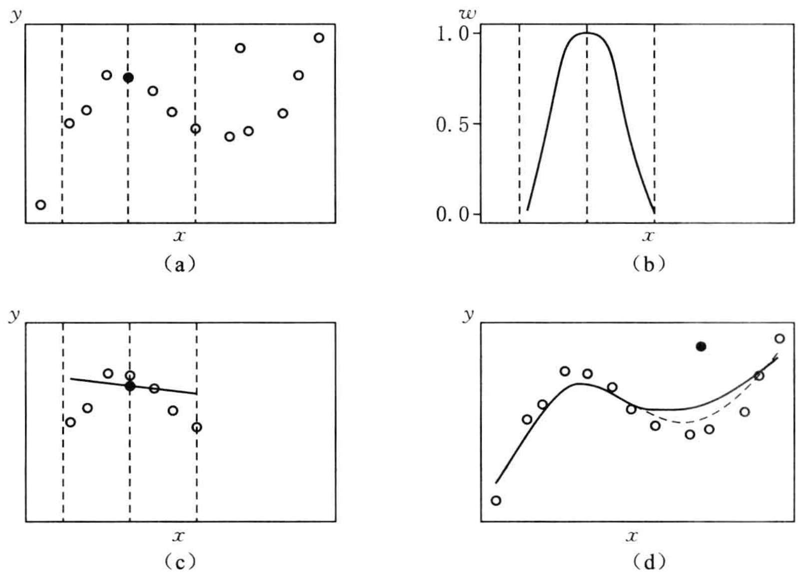
对局部权重散点图平滑法的简写为 lowess(Cleveland，1985)，它可以帮助对每个 $x_i$ 获得对应的平滑拟合值 $\hat{y}_i$(其中 $y$ 和 $x$ 是散点图中的垂直与水平变量)。为了获得平滑值，lowess 步骤会针对每个观测 $i$ 拟合数据的 $n$ 条回归线，并突出接近 $x_i$ 的 $x$ 值。图 A. 1 显示了 lowess 步骤。由于 lowess 是需要精密计算的，因此需要一个特殊的电脑程序来运行，但是这个程序很容易写，且越来越普遍。

选择一个平滑分量：选择一个数据中的分量 $0 < f \leqslant 1$，使得每个对应于 $r = [fn]$ 的数据值的拟合都包含在内，其中中括号代表取其最接近的正数。通常 $f = 1/2$ 或 $2/3$ 较为适用。较大的 $f$ 值将产生更平滑的结果。

局部权重回归：对每个 $x_i$，选择最接近 $x$ 的 $r$ 值，用 $x_1^{(i)}$，…，$x_r^{(i)}$ 表示，参见图 A. 1(a)。对这个观测的窗口一半的宽度即为到最远的 $x_j^{(i)}$ 的距离，即 $W_i = | x_i - x_r^{(i)} |$。对窗口内 $r$ 个观测中的每一个，计算权重 $w_j^{(i)} = w_t[(x_j^i - x_i)/W_i]$，其中 $w_t$ 是三次方的权重函数。

$$w_{t(z)}=\begin{cases}0 & (|z|\geqslant 1)\\ (1-|z|^3)^3 & (|z|<1)\end{cases}$$

在这里，$z$ 仅代表三次方函数的自变数，即 $(x_j^{(i)}-x_i)/W_i$。因此当 $x_j^{(i)}$ 接近窗口的界线（且最大的为 $x_i$）时，$w_j^{(i)}$ 减小至 0。




注：(a)以 $x_5$ 为中心建立一个窗口，以包含数据的 $f=1/2$，因此 $r=[fn]=7$ 个点在窗口中。点$(x_5, y_5)$用一个实心点表示。(b)三次方权重函数在窗口的界线处下降至 0，且当时 $x=x_5$ 时取得最大值。(c)对窗口内的 7 个观测进行局部回归，并使用由(b)中权重函数获得的权重。$x_5$ 处的 lowess 拟合值 $\hat{y}_5$是用实心点表示的。每个观测都重复步骤(a)、(b)和(c)以获得整个数据集的全部 14 个拟合值。(d)通过连接这些拟合值 $\hat{y}$，…，$\hat{y}_{14}$，就能获得 lowess 曲线（实线）。注意，这条曲线被拖向下方的观测（拟合点）。这条虚线显示了对奇异值赋予低权重是如何帮助获得更稳健的拟合（连接拟合值 $\hat{y}'_1$，…，$\hat{y}'_{14}$）的。

资料来源：《基本数据绘图》(*The Element of Graphing Data*)，W. S. Cleveland。版权© 1985 贝尔电话实验室，Murray Hill，NJ。获得了 Wadsworth 和 Brooks/Cole 的高级图书和软件的同意，Pacific Grove，CA 93950。

**图 A.1 lowess 是如何运作的**

参见图 A.1(b)，则拟合这个回归方程：

$$y_j^{(i)} = a_i + b_i x_j^{(i)} + e_j^{(i)}$$

将使 $\sum_{j=1}^{r} w_j^{(1)} e_j^{(i)2}$ 最小(参见附录 8 的权重最小二乘回归)。计算拟合值 $\hat{y}_i = a_i + b_i x_i$ 。注意,对每个 $i = 1, \cdots, n$ 都有一个拟合方程,且都有一个拟合的值。

对奇异值赋予低权重:计算残差 $e_i = y_i - \hat{y}_i$。计算能够低估具有较大残差观测的稳健权重:$\delta_i = w_b(e_i/6M)$,其中 $M$ 是残差 $|e_i|$ 绝对值的中位数,$w_b$ 是二次方权重函数:

$$w_{b(z)} = \begin{cases} 0 & (|z| \geqslant 1) \\ (1-z^2)^2 & (|z| < 1) \end{cases}$$

稳健局部权重回归:重复局部权重回归,但是在单独回归中,使用混合权重 $\delta_j w_j^{(i)}$ 得出新的拟合值 $\hat{y}'_i$。

## 附录 8 权重最小二乘估计

假设回归模型为:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$
$$\varepsilon_i \sim \text{NID}(0, \sigma^2) \qquad [\text{A. 4}]$$

误差的标准差与 $x_1$ 是成比例的,$\sigma_i = \sigma x_{1i}$。在方程 A. 4 两边都除以 $x_{1i}$ 可以获得:

$$\frac{y_i}{x_{1i}} = \beta_0 \frac{1}{x_{1i}} + \beta_1 + \beta_2 \frac{x_{2i}}{x_{1i}} + \cdots + \beta_k \frac{x_{ki}}{x_{1i}} + \frac{\varepsilon}{x_{1i}} \qquad [\text{A. 5}]$$

且由于 $x_{1i} = \sigma_i/\sigma$,则最后一项变为 $\varepsilon'_i = \sigma\varepsilon_i/\sigma_i$。因为 $V(\varepsilon'_i) = \sigma^2 V(\varepsilon_i)/\sigma_i^2 = \sigma^2$ 是恒定的,所以对方程 A. 5 的转换可以通过最小二乘回归获得,包括对 $y_i/x_{1i}$ 根据一个恒定的回归因子 $1/x_{1i}$ 和对 $x_{2i}/x_{1i}$ 根据 $x_{ki}/x_{1i}$ 进行的回归,这样可以获得对 $\beta$ 的估计和它们的标准误。这个步骤与将权重的平方和 $\sum e_i^2/\sigma_i^2$ 最小化是

等价的，这可以得到方程 A. 4 的最大似然估计。只要误差方差是一个一致的百分常量 $V(\varepsilon_i) = \sigma^2 w_i$，这种方法就是有效的(参见 Weisberg，1985:第 4 章)。

## 附录 9　矫正异方差的最小二乘标准误

回顾附录 1，最小二乘估计量的协方差矩阵为：

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad [\text{A. 6}]$$

在误差方差一致性的假设下，有 $V(\mathbf{y}) = \sigma^2\mathbf{I}$，方程 A. 6 将其简化为一般的形式，$V(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$。如果误差是异方差但是独立的，则 $V(\mathbf{y}) = \sum = \text{diag}(\sigma_1^2, \cdots, \sigma_n^2)$，且

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sum\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

因为 $E(\varepsilon_i) = 0$，第 $i$ 个误差方差位 $\sigma_i^2 = E(\varepsilon_i^2)$，这表明了估计 $V(\mathbf{b})$ 的可能性

$$\widetilde{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\sum}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad [\text{A. 7}]$$

其中 $\hat{\sum} = \text{diag}(e_i^2, \cdots, e_n^2)$，$e_i$ 是观测 $i$ 的最小二乘残差。White(1980)证明了方程 A. 7 是 $V(\mathbf{b})$ 的一致估计量。

例如，对 Ornstein 的连锁董事会数据，White 的方法获得的估计标准误与一般方程获得的结果基本类似(在表 6. 1 中给出)。事实上，对大部分系数而言，矫正的误差方差比未矫正的要小一些。但是，公司资产平方差的系数的矫正误差方差为 0. 028，比未矫正的误差方差 0. 019 大了许多。

## 附录 10　当误差方差不一致时最小二乘估计的有效性和可信性

不一致误差方差对最小二乘估计有效性的影响和对最小二乘推论可信性的影响基于许多因素，包括样本规模、$\sigma_i^2$ 的变异程度、$x$ 值的模式和误差方差与 $x$ 之间的关系。因此，我们无法获得一个普遍使用的结论，但是下面的简单例子

可以说明很多问题且支持本书所给的建议。

假设 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$，其中 $\varepsilon_i \sim \text{NID}(0, \sigma_i^2)$ 且 $\sigma_i = \sigma x_i$（与附录 8 中的一样）。则 OLS 的估计量 $b_1$ 没有 WLS 估计量 $\hat{\beta}_1$ 有效，因为后者的情况是 $\beta_1$ 的最大有效无偏估计量。

抽样方差 $b_1$ 与 $\hat{\beta}_1$ 的公式很好推导(例如 Kmenta，1986：第 8 章)。OLS 估计量的有效性与 WLS 最佳的估计量有效性的比较可以通过 $V(\hat{\beta}_1)/V(b_1)$ 得出，而 OLS 的相对精确度为这个比例的平方根，即 $\text{SE}(\hat{\beta}_1)/\text{SE}(b_1)$。

现在，假设 $x$ 是在区间$[x_0, \alpha x_0]$上一致分布的，其中 $x_0 > 0$ 且 $\alpha > 0$，则 $\alpha$ 是 $x$ 的最大值与最小值的比例(因此也就是最大与最小的 $\sigma_i$)。OLS 估计量的相对精度随着样本规模的增大而逐渐变得稳定，且当 $\alpha = 2$ 时超过 90%，$\alpha = 3$ 时超过 85%，即便 $n$ 很小，只有 20。对 $\alpha = 10$，使用 OLS 的代价则较大，但是当 $n \geqslant 20$ 时，相对精度仍然超过 65%。

基于最小二乘估计的统计推断有效性，对一般模式的不一致误差方差并不十分敏感。这里，我们需要比较一般估计量 $V(b_1)$的期望与真正的抽样方差 $b_1$。同样，$E[\hat{V}(b_1)]$的公式很容易推导出来(参见 Kmenta，1986：第 8 章)。$E[\hat{V}(b_1)]/V(b_1)$的平方差显示了相对标准误项的结果。例如，针对 $n \geqslant 20$ 的情况，当比例为 98%时，$\alpha = 2$；当比例为 97%时，$\alpha = 3$；当比例为 93%时，$\alpha = 10$。

# 参考文献


Anscombe, F. J. (1960) "Rejection of outliers" [with commentary]. *Technometrics* 2:123—166.

Anscombe, F. J. (1961) "Examination of residuals." *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability* 1:1—36.

Anscombe, F. J. (1973) "Graphs in statistical analysis." *American Statistician* 27:17—22.

Anscombe, F. J., and Tukey, J. W. (1963) "The examination and analysis of residuals." *Techonometrics* 5:141—160.

Atkinson, A. C. (1985) *Plots, Transformations, and Regression: An introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford: Clarendon.

Bartlett, M. S. (1937) "Properties of sufficiency and statistical tests." *Proceedings of the Royal Society* A 160:268—282.

Beckman, R. J., and Cook, R. D. (1983) "Outliers." *Technometrics* 25:119—163.

Belsley, D. A. (1984) "Demeaning condition diagnostics through centering" [with commentary]. *American Statistician* 38:73—93.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.

Box, G. E. P., and Cox, D. R. (1964) "An analysis of transformations." *Journal of the Royal Statistical Society, Series* B 26:211—252.

Breusch, T. S., and Pagan, A. R. (1979) "A simple test for heteroscedasticity and random coefficient variation." *Econometrica* 47:1287—1294.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.

Chatterjee, S., and Hadi, A. S. (1988) *Sensitivity Analysis in Linear Regression*. New York: John Wiley.

Chatterjee, S., and Price, B. (1977) *Regression Analysis by Example*. New York: John Wiley.

Cleveland, W. S. (1985) *The Elements of Graphing Data*. Belmont. CA: Wadsworth.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981) "A comparative study of tests for homogeneity of variance, with applications to the outer continental shelf bidding data." *Techometircs* 23:351—361.

Cook, R. D. (1977) "Detection of influential observation in linear regression." *Technometrics* 19:15—18.

Cook, R. D., and Weisberg, S. (1982a) "Criticism in regression," in S. Leinhardt (ed.) *Sociological Methodology* (pp. 316—361). San Francisco: Jossey-Bass.

Cook, R. D., and Weisberg, S. (1982b) *Residuals and Influence in Regression*. London: Chapman and Hall.

Cook, R. D., and Weisberg, S. (1983) "Diagnostics for heteroscedasticity in regression." *Biometrika* 70:1—10.

Daniel, C., and Wood, F. S. (1980) *Fitting Equations to Data* (2nd ed.). New York: John Wiley.

Davis, C. (1990) "Body image and weight preoccupation: A comparison between exercising and non-exercising women." *Appetite* 15:13—21.

Diaconis, P., and Feron, B. (1983) "Computer intensive methods in statistics." *Scientific American* 248(5):116—130.

Draper, N. R., and Smith, H. (1983) *Applied Regression Analysis* (2nd ed.). New York: John Wiley.

Duncan, O. D. (1961) "A socioeconomic index for all occupations," in A. J. Reiss, Jr., with O. D. Duncan, P. K. Hatt, and C. C. North, *Occupations and Social Status* (pp. 109—138). New York: Free Press.

Ericksen, E. P., Kadane, J. B., and Tukey, J. W. (1989) "Adjusting the 1980 Census of Population and Housing." *Journal of the American Statistical Association* 84:927—944.

Fox, J. (1984) *Linear Statistical Models and Related Methods*. New York: John Wiley.

Fox, J. (1990) "Describing univariate distributions," in J. Fox and J. S. Long(eds.) *Modern Methods of Data Analysis* (pp. 58—125). Newbury Park, CA:Sage.

Fox, J., and Monette, G. (in press) "Generalized collinearity diagnostics." *Journal of The American Statistical Association*.

Fox, J., and Suschnigg, C. (1989) "A note on gender and the prestige of occupations." *Canadian Journal of Sociology* 14:353—360.

Gallant, A. R. (1975) "Nonlinear regression." *American Statistician* 29:73—81.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (eds.) (1985) *Exploring Data Tables, Trends, and Shapes*. New York: John Wiley.

Hoaglin, D. C., and Welsch, R. E. (1978) "The hat matrix in regression and ANOVA." *American Statistician* 32:17—22.

Hoerl, A. E., and Kennard, R. W. (1970a) "Ridge regression: Applications to nonorthogonal problems." *Technometrics* 12:69—82.

Hoerl, A. E., and Kennard, R. W. (1970b) "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics* 12:55—67.

Kish, L. (1965) *Survey Sampling*. New York: John Wiley.

Kmenta, J. (1986) *Elements of Econometrics* (2nd ed.). New York: Mascmillan.

Leamer, E. E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: John Wiley.

Little, R. J. A., and Rubin, D. B. (1990) "The analysis of social science data with missing values," in J. Fox and J. S. Long (eds.) *Modern Methods of Data Analysis* (pp. 374—409). Newbury Park, CA: Sage.

Mallows, C. L. (1973) "Some comments on $C_p$." *Technometrics* 15:661—676.

Mallows, C. L. (1986) "Augmented partial residuals." *Technometrics* 28:313—319.

Monette, G. (1990) "Geometry of multiple regression and interactive 3-D graphics," in J. Fox and J. S. Long (eds.) *Modern Methods of Data Analysis* (pp. 209—256). Newbury Park, CA: Sage.

Mosteller, F., and Tukey, J. W. (1977) *Data Analysis and Regression: A Second Course in Sta-*

*tistics*. Reading, MA: Addison-Wesley.

National Opinion Research Center. (1989) *General Social Survey* [data set]. Chicago: Author.

Ornstein, M. D. (1976) "The boards and executives of the largest Canadian corporations: Size, composition, and interlocks." *Canadian Journal of Sociology* 1:411—437.

Ostrom, C. W., Jr. (1990) *Time Series Analysis: Regression Techniques* (2nd ed.). Newbury Park, CA: Sage.

Pineo, P. C., and Porter, J. (1967) "Occupational prestige in Canada." *Canadian Review of Sociology and Anthropology* 4:24—40.

Rousseeuw, P. J., and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. New York: John Wiley.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Statistics Canada. (1971) *Census of Canada* (Vol. 3). Ottowa, Canada: Author.

Stine, R. (1990) "An introduction to bootstrap methods: Examples and ideas," in J. Fox and J. S. Long (eds.) *Modern Methods of Data Analysis* (pp. 325—373). Newbury Park, CA: Sage.

Theil, H. (1971) *Principles of Econometrics*. New York: John Wiley.

Tobin, J. (1958) "Estimation of relationship for limited dependent variables." *Econometrica* 26: 24—36

Tukey, J. W. (1977) *Exploratory Data Analysis. Reading*, MA: Addison-Wesley.

Velleman, P. F., and Hoaglin, D. C. (1981) *Applications, Basics and Computing of Exploratory Data Analysis*. Boston: Duxbury.

Velleman, P. F., and Welsch, R. E. (1981) "Efficient computing of regression diagnostics." *American Statistician* 35:234—241.

Weisberg, S. (1985) *Applied Linear Regression* (2nd ed.). New York: John Wiley.

White, H. (1980) "A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity." *Econometrica* 38:817—838.

Wonnacott, T. H., and Wonnacott, R. J. (1990) *Introductory Statistics* (5th ed.). New York: John Wiley.

# 译名对照表

| | |
|---|---|
| added-variable plot | 添加变量散点图 |
| autocorrelated error | 误差自相关 |
| biased estimation | 有偏估计 |
| bulging rule | 撑压法则 |
| component-plus-residuals plots | 分量-残差散点图 |
| conditioning | 调节 |
| cutoffs | 截断点 |
| deleted studentized residual | 删除学生残差 |
| diagnostics | 诊断 |
| elasticity | 弹性 |
| externally studentized residual | 外部学生残差 |
| general linear model | 一般线性模型 |
| hat matrix | 预测矩阵 |
| hat-values | 预测值 |
| high-breakdown estimator | 高分项估计量 |
| influential observation | 强影响观测值 |
| locally weighted scatterplot smoother | 局部权重散点图平滑法 |
| mean-squared error | 均方差 |
| mean-squared error | 分层 |
| model respecification | 模型的重新确定 |
| multicollinearity | 多元共线性 |
| multiple mode | 多路方式 |
| multiple correlation | 复相关系数 |
| multiple regression | 多元回归 |
| non-normality | 非正态 |
| normal quantile-comparison plot of residuals | 残差的正态分位数比较散点图 |
| onstructed variables | 构造变量 |
| outliers | 奇异值 |
| overfitting | 过度拟合 |
| partial effect | 局部效应 |
| partial-regression leverage plots | 偏回归影响力散点图 |
| partial-regression plots | 偏回归图 |
| power transformation | 次方转换 |
| ridge constant | 岭常数 |
| ridge regression | 岭回归 |

| | |
|---|---|
| robust estimator | 稳健估计量 |
| robust regression | 稳健回归 |
| score tests | 计分检验 |
| smoothed histogram | 平滑直方图 |
| standard deviation | 标准差 |
| standardized residual | 标准化残差 |
| stem-and-leaf display | 茎叶图 |
| studentized residuals | 学生残差 |
| time-series regression | 时间序列回归 |
| truncation or censoring | 截断或者删节 |
| unit-normal deviate | 单位正态变异 |
| variance-inflation factor | 方差膨胀因子 |

# 虚拟变量回归

## 作者简介

**梅丽莎·A. 海蒂(Mellissa A. Hardy)**

美国佛罗里达州立大学社会学院教授和老龄化研究中心主任。目前的研究领域包括大龄男女劳动力参与问题、退休政策和实施的变化问题、社会流动问题及贫困妇女化研究。

## 译者简介

**贺光烨**

2007 年获南京大学理学学士学位,目前是香港科技大学社会科学部硕士研究生。

# 序

第一次听到"虚拟变量"这个词的时候，许多定量研究方法的学生都会觉得有趣，但很快他们就会意识到，这个听上去"虚拟"的方法，在定量研究中却起着至关重要的作用。我们知道，在回归分析中，用定序或者名义变量作为自变量来进行回归分析，既不能有效地反映因变量与自变量之间的实际关系，而且又容易出现拟合不足的情况。然而，引入了"虚拟变量"的概念后，我们就可以在不违反测量相关假设的情况下，运用最小二乘法进行回归分析。

那到底什么是"虚拟变量"呢？简单地说，虚拟变量是由原先的定性变量构建出来的二分变量。对于二分法，通常需要 $G-1$ 个数字来涵盖所有信息，其中 $G$ 为原先类别的个数。例如，在民意调查中，如果我们希望表达公民的政治兴趣（其中包括 3 个类别——非常同意、有点同意、不同意），研究者必须构建两个二分变量。假设它们分别为 $X_1$（编码 1 表示非常同意，0 表示除非常同意外的类别）和 $X_2$（编码 1 表示有点同意，0 表示除有点同意外的类别），如果 $X_1$、$X_2$ 两个变量的编码都为 0，那么暗示了受访者所属类别为不同意。在这里，"不同意"这个类别被设置成了底线，或者说是一个参照组，从而 $X_1$ 和 $X_2$ 的回归系数都是在其他组与该组比较后估计得到的。

但是为什么选择"不同意"作为参照组而不选其他类别，如"有点同意"呢？曾经使用过虚拟变量的研究者基本都遇到过这样的问题。在这里，Hardy 教授给出了明确的答案。在本书中，一个有关收入的、精心设计的例子贯穿全文，从一个简单模型（含有一个虚拟变量的回归模型，常常将其简化到均值差异的检验）到一系列复杂模型（含有多个虚拟变量、多个定量变量及多个交互项的回归模型）。所幸的是，通过严谨的语言叙述，这种复杂性可以用不同条件下所得的回归系数来表达。

对虚拟变量回归有了基本了解后，Hardy 教授还提出了有关虚拟变量回归的一些特殊问题。除此以外，她还对如何处理异方差性，在因变量取对数或者 logit 后，如何对回归系数进行诠释，如何在显著性检验下进行多重比较，如何进行效果编码和对比编码以及如何检验曲线性和如何进行分段线性回归作出了解释。

总之,本书以通俗易懂的语言,从不同角度对虚拟变量的用法进行了详述。在有关统计方法的书籍中,没有任何一个作者可以如此全面地诠释一个问题。可以说,这本书无疑是一部有关虚拟变量回归的重要著作。

迈克尔·S.刘易斯-贝克

# 第 1 章 | **简介**

回归分析是定量分析里运用最灵活、最广泛的一种方法。一个典型的回归模型试图将因变量 $Y_i$ 映射到一系列特定的自变量 $X_i$ 上，并通过相应的线性函数来解释因变量 $Y_i$ 的变异。利用最小二乘估计，我们可以得到一个预测方程，用来估计自变量的条件均值，即特定自变量组合下的 $Y$ 的期望值，从而得到因变量的条件均值。当自变量像定量变量那样可测量时，我们可以假设其为一系列任意的相对零点且间隔大致相等的定量变量，此时，所有可能的 $Y$ 的期望值都是无限的。此外，当因变量和自变量都是定量变量时，其相应的关系可用几何图形表示。

在二元回归中，我们预测 $Y$ 为唯一自变量的函数，则两个变量之间的关系可由回归线直接表示。线上所有的点代表 $Y$ 的条件均值。当有第二个自变量包含到函数中时，一维回归线扩展成二维，一个由南北方向和东西方向的线组成的平面生成了，此时代表 $Y$ 的条件均值的是所有处于该平面上的点。由此可见，当自变量的数量增加时，这些原则是保持不变的，尽管其几何形态可能变得难以描述。

但如果所有用来预测的自变量都用间隔尺度来衡量，那么回归模型的有效性将会受到严重制约。我们研究的问题经常涉及组差异，如社会学家感兴趣的对民族/种族差异、性别差异，或行为、态度及社会经济特征的区域差异的解释。又如，市场调研人员希望从人口统计数据中了解消费者偏好。研究人员常常想知道对于所有组别，自变量的影响是否一样，或者在同一关系的强度或方向上，组差异是否依然存在。由此可知，我们大多数的研究问题是为了区别各级因变量下的组差异以及不同自变量影响下的因变量的组差异。

当感兴趣的自变量为定性变量时(即“只在名义水平上测量”),我们需要一种方法,它既能定量地代表这种信息,又能防止将不切实际的测量假设强加于分类变量。例如,我们可以将职业分类按 1 到 12 进行编码(该分类用于人口普查中的单数代码),但我们不可以简单地说,职业的范围是从低值 1 到高值 12,因为这种描述是建立在假定的间隔相等的基本衡量标准上的。定义一系列虚拟变量可以使我们捕捉到分类方案里的分级信息,然后把此信息用到标准回归估计中。事实上,回归方程中的自变量可以是任意定性和定量预测因子的组合。

例如,“社会资源是通过收入进行分配的”,这个现象既是那些对不平等感兴趣的学者所关注的焦点,也是那些努力为维持生活水平而奋斗的人民群众所关心的问题。我们关于社会公正的信念往往建立在对资源分布的认识上,以及是否有某些特定团体在分配过程中处于优势或劣势。我们知道,对于研究劳动收入分配中的歧视,有一种常见的方法,即首先确定一个组差异,比如男人和女人的差异或者黑人和白人的区别,以这个组差异作为在劣势群体的总效应,然后探讨加入其他决定性因素后,这个总差异如何变化,它是不是仍然维持不变?通过此方法,那些形成于社会进程中的、可察觉的不平等从而可被识别。

为了之后讨论统计方法时的连贯性,我会引用一个例子,即预测收入是个体特征的函数,并用定性或定量变量描述相应的个体特征。我所用的数据来自于全国老年男性纵向调查。通过第一次入户结果[1]可知,在最初的样本中,我们的研究对象大约为美国 1500 万 45 岁至 59 岁且未收容到专门机构(如监狱、精神病院)的男性。在该例中,我们比较感兴趣的变量包括种族、职业(美国人口普查分类)、教育(受教育年限)和工作任期(在同一个雇主下的工作年限)。尽管其他变量,例如劳动力的供给、工作技能、健康等也可以被假设为(通过薪酬得到的)年收入的预测因子,但是对于此例,我们不予考虑,而用只含有 4 个预测因子的函数提供一个定性定量相结合的估测。通过讨论逐步复杂化的模型来阐述虚拟变量回归的方法,我会尽量解释清楚有关任意特定的虚拟变量的系数是如何随模型整体而变的问题。同时,我还希望通过这些努力,减少读者在不适用的情况下,对此方法进行演绎的可能性。

本书以讨论我们最初关注的问题——黑人和白人之间的收入差异(用"美元/年"衡量)开始。之后,我们会不断加入新的假设并逐步建立复杂的模型进行检验。我们所要估计的是,当控制了更多的自变量(包括定性的或定量的)后,黑人和白人之间的平均收入差异是否仍然存在。还有,各个自变量的净效应在黑人和白人中是否一样。最后,我们将使用虚拟变量回归的形式来估计种族对回归模型所有参数的具体影响。有关这个逐步深入的过程,我们将在第4章具体描述。尽管未必所有读者都对收入分配这个话题感兴趣,但是由于其中所涉及的方法比较直接简单,所以适合各个学科背景的读者。此外,这些对模型的解释方式很容易扩展到其他实质性研究里,所以我还是选择了这个例子。本书第5章提供了一个有关对虚拟变量的替代性编码策略的简短描述。在第6章中,我们会把视线从单个问题上移开,而更加关注虚拟变量在其他研究里的运用。

## 多元线性回归回顾

随后的讨论均建立在有关单次方程回归模型的概念、偏相关的方法以及假设检验的基础上,这些读者都已比较熟悉。如果读者对这些还不太清楚,建议先阅读有关回归的介绍性书籍,其中有较早的卷本(比如,Berry & Feldman, 1985; LewisBeck, 1980; Schroeder, Sjoquist & Stephan, 1986),还有一些基本的统计书籍(比如,Bohrnstedt & Knoke, 1982; Cohen & Cohen, 1983),这些书可为我们即将讨论的问题提供非常有用的信息。

在文中,我们还会用到一些符号语言,因此,现在来回顾一些基本的符号。假设我们有一个定量的因变量($Y_i$),其为3个定量自变量 $X_{1i}$、$X_{2i}$、$X_{3i}$ 的线性函数,则总的回归函数可写为:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i = \beta_0 + \sum \beta_k X_{ki} + u_i \qquad [1.1]$$

其中,$k$ 为第 $k$ 个自变量,$i$ 为第 $i$ 个观测值。该方程表达了 $Y_i$ 是 $X_{1i}$、$X_{2i}$、$X_{3i}$ 以及随机误差项 $u_i$ 的线性函数;$\beta_0$ 是截距项,其可解释为当所有自变量均为0

时，$Y_i$ 的值；$\beta_1$ 为总体偏回归系数，表示当控制了其他自变量后，$X_{1i}$ 每变化一个单元，$Y_i$ 增加或者减少的量；$\beta_2$ 和 $\beta_3$ 同样也是总体偏回归系数，其分别是变量 $X_{2i}$ 和 $X_{3i}$ 的系数。由此可知，总体回归函数对于给定的自变量 $X_{ki}$，提供了 $Y_i$ 的条件均值或期望值。因此，我们可以通过样本回归函数，用最小二乘估计来估测其所在的总体参数。

$$Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + B_3 X_{3i} + e_i \tag{1.2}$$

每个回归系数 $B_0$、$B_1$、$B_2$、$B_3$ 既是方程 1.1 中相对应的总体参数的点估计，也是统计抽样分布的观测值。我们用 $e_i^2$ 的观测值来估计总体方差和抽样分布里 $B_0$、$B_1$、$B_2$ 和 $B_3$ 的标准误，从而可以评估所得出的估测的显著性意义，进而对 $Y_i$ 和 $X_k$ 的关系作出结论。此外，标准差还可以用来构造区间估计，该区间通常被称为"置信区间"，其对评估有关假定的统计证据很有用。当如下假设都成立时，我们就可以用最小二乘法来分析这些样本数据了：

(1) $E(u_i \mid X_k) = 0$；即，在给定的 $X_k$ 值下，$u_i$ 的平均值为 0。

(2) $\text{cov}(u_i, u_j) = 0$；即，干扰项对所有 $i \neq j$，都是独立的。

(3) $\text{var}(u_i) = \sigma^2$；即，对任意 $X_k$ 的取值，$u_i$ 的方差都是非负常数 $\sigma^2$，这也是同方差性的假设。

(4) $\text{cov}(u_i, X_k) = 0$；即，干扰项和解释变量都是独立分布的，彼此不相关。

在这些假设下，OLS 估计是最好的无偏估计，"最好"是因为在所有线性无偏估计中，其方差最小。

异方差性的问题通常与截面数据(描述整体的单元群在一个特定时间点的数据)相关，自相关作用常常与时间序列数据(描述实体在一段时间内的数据)有关。而虚拟变量在研究截面数据和时间序列数据里，都扮演着非常重要的角色。在截面数据中，虚拟变量可以用来估计各群体之间的差异或者加入某一群体后，是否会改变其他解释变量的效应问题。同样，在时间序列数据中，虚拟变

量可以用来确定两个时间段是否有区别，或检验不同时间段上，其解释变量所造成的影响的稳定性如何(Gujarati, 1970)。由于虚拟变量通常既可以在截面数据分析也可以在时间序列分析中定义观测组，因此，研究人员必须谨慎处理这两种情况下的异方差问题。在截面分析中，我们可以通过指定虚拟变量以获得组差异，但是，其前提是那些潜在的异质组的信息已经被合并了。如果这些群体的误差方差显著不同(即，如果我们违反了同方差性的假设)，那么单个回归系数的显著性检验将会变得很不可靠。像这样类似的问题也可能在时间序列模型分析中出现，因为虚拟变量常用于检测两个或多个时间段的系数的稳定性，所以，如果误差方差在不同时间段中呈现出显著区别，那么异方差会使回归检验非常有争议(Maddala, 1992)。有关假设的讨论、违反假设的种种后果的详述以及有关处理这些违反假设的种种补救方法，在很多中级统计教材里都有提及。因这些均与虚拟变量的使用有关，所以我也会在本书中予以讨论，有关异方差性及自相关作用的问题，我们会在后文详述。

通过 $R^2$，即多重相关系数的平方，我们可以评估回归模型对样本数据的整体拟合度。$t$ 检验通常用来检验单个回归系数的统计显著性，为什么我们用 $t$ 分布而不是 $z$ 分布呢？其原因在于，我们一般不知道总体方差 $\sigma^2$ 的值。因此，我们只能用样本的误差方差作为对总体方差的估计。当检验零假设，即其效应或局部效应是否等于 0 时，$t$ 检验可以把参数估计的比率降低到其标准误。

由于 3 个自变量均包含在统计规范中，$B_1$、$B_2$、$B_3$ 可用来估计 $X_1$、$X_2$、$X_3$ 对 $Y$ 的局部效应。一般来说，局部效应不等于当 $Y$ 只对一个自变量回归时产生的二元效应，因为在一个给定的规范中，自变量通常相互关联或与 $Y_i$ 共协方差。当其中一个自变量(例如，$X_{1i}$)与其他一个或多个自变量完全相关(即，该自变量是其他一个或多个自变量的线性函数)时，那么该样本估计是不确定的。直观地讲，我们可以把这种不确定归因于缺乏“唯一的”信息：“$X_{1i}$ 分布中的信息直接照搬方程的右边所包含的统计信息，当 $X_{1i}$ 没有提供任何净分布信息(例如，局部信息或唯一信息)时，我们不可能估计 $X_{1i}$ 对 $Y$ 的净效应(不论局部或者唯一)。”这就是完美的多重共线性。从统计学上讲，当估测偏回归系数时，在我们可以明确“其他自变量被控制了”的意义后，该不确定性就可以被解

释了。在统计学上，“保持不变”需要我们将与模型中其他自变量有关的变异从 $Y_i$ 的分布中移除。由此可见，统计上的“控制”是一个分割变异的过程。在我们的样本回归函数中（方程 1. 2），当确定 $B_1$ 时，我们会移除 $X_{2i}$ 和 $X_{3i}$ 对 $Y$ 的影响。换句话说，当 $X_{2i}$ 和 $X_{3i}$ 不能在样本中变化时，我们就看不到那些与 $X_{2i}$ 和 $X_{3i}$ 的变异相关的部分体现在 $Y$ 或 $X_{1i}$ 的分布里。因此，在估计 $X_{1i}$ 对 $Y$ 的局部效应时，我们不可以把那部分变异，即与 $X_{2i}$ 和 $X_{3i}$ 的变异相关的部分变异考虑在内。从本质上来讲，$X_{1i}$ 对 $Y_i$ 的局部效应是基于两个残差分布——移除了 $X_{2i}$ 和 $X_{3i}$ 对 $Y_i$ 的线性效应后的 $Y_i$ 的残差分布，以及移除了 $X_{2i}$ 和 $X_{3i}$ 对 $X_{1i}$ 的线性效应后的 $X_{1i}$ 的残差分布。当存在完美的共线性时，$X_{1i}$ 的残差分布是一个常数——0。

当我们把虚拟变量加入回归方程中后，回归估计的逻辑是不变的，即我们将根据因变量来预测条件均值，也就是说，通过把给定数值的自变量代入方程而得出 $Y$ 的均值。其区别在于，虚拟变量的编码通常代表每个不同组群，或者根据是否有某个或某些特征，将虚拟变量分成有或者没有两种状态。因此，预测一个虚拟变量编码的特定组合的 $Y$ 的期望值与预测组群的均值没有什么差别。如果是这样，当自变量是连续的时候，该偏相关的程序对解释虚拟变量就至关重要。

# 第 2 章 | 构建虚拟变量

分类数据的编码要求详尽且彼此不相关的分类全部被列出来。该原则同样适用于虚拟变量的创建。我们需要构造一个足够庞大的虚拟变量集合，从而将原先在定性范围内的所有信息都尽量表达出来。分类变量可以是两分或者多分的。一个有 $j$ 个类别的分类变量需要 $j-1$ 个虚拟变量来获得初始差异集合中的所有分布信息。因虚拟变量通常为二分变量，所以我们通常用二进制编码(0，1)进行区分。所有在特定类别中的成员会被分配到代码 1 中，其他不在该类别的成员被分配到代码 0 中。根据这个编码原则，我们为一个给定的大类构造了一系列虚拟变量，原因在于在真实数据中，只有特定的受访者会被编码为 1，而且对任意的受访者，每个受访者只可对应大类中一个且只有一个的虚拟变量。我们可以把二进制编码想象成电气开关：编码 1 亮起时，表示一个给定类对一个受访者信息(例如，他/她是某一特定组群的成员或者他/她具有某一特定的特征)"开启"。对那些非成员来说，虚拟变量会切换到"关闭"状态(表示某些特征不存在)。

我们常用 $j-1$ 个虚拟变量来描述一个包含 $j$ 个类别的定性变量，这样做的原因在于其直接符合古典线性回归的模型，尤其在自变量中，没有完全共线性的假设要求。在一个模型里，任何解释变量都不可能与其他解释变量存在完美共线关系。那么，假设在我们的例子中，用虚拟变量来表示种族。被编码成 1 的虚拟变量(黑人)代表非洲裔美国人，若我们再加入第二个虚拟变量(白人)也编码为 1，使其代表那些非非洲裔美国人，则在该模型中，就已经构造了两个自变量之间的完全线性关系，因为 BLACK $=1-$ WHITE。所以，白人中的信息对估计是多余且不必要的。

当原先的变量如之前提到的种族(黑人、白人)只有两类时,只需一个单一的虚拟变量便足以捕捉全部信息了。其中,没有用虚拟变量命名所表示的类为参照组。如果原先的变量有两个以上的类,虚拟变量的数目就取决于在分析中需要比较的不同类的数目。我们以职业为例,在普查里,其通常以 12 位数编码来衡量。在该例中,我们以第 12 类作为参照组,那么,最多可以创建 11 个虚拟变量。为了解释得更清楚、更容易理解,我们将排除农场管理人员和雇农,并把剩余的组归为一类,因此,我们只需考虑 6 个类别:$OCC_1$(高级白领,如专家、经理等)、$OCC_2$(初级白领,如文员、推销员等)、$OCC_3$(技术工人,如木匠、水管工人、电工等)、$OCC_4$(操作工人,如焊工、织布工人及在生产制造中的装订工人)、$OCC_5$(非家庭服务工作者,如理发师、守卫、实习护士等)及 $OCC_6$(劳工,如渔民、锯木工人、货车司机等)。根据前述的规则可知,这组含有 6 个详尽且互斥的类别需要 5 个虚拟变量来表示原先定性变量的所有信息,其中,5 个虚拟变量分别表示不同的类,剩下的第 6 类,即被排除的类(未用虚拟变量命名所表示的类)为参照组。

## 选择参照组

在为一些分类变量编码时,我们必须选择参照组,即我们想把哪个或哪些组作为要比较的类。对每个分类变量,我们必须指定一个类作为参照组。在我们的例子中,如果选择“白人”作为参照组,那么虚拟变量“黑人”在二元回归里的系数就代表美国黑人相对于白人的平均收入是多少。换句话说,回归系数表达了黑人与白人之间的平均收入差异。那么如果以“白人”为参照组,则虚拟变量“黑人”在二元回归中的系数 $B$ 可表示为如下:

$$B_{\text{BLACK}} = \bar{Y}_{\text{BLACK}} - \bar{Y}_{\text{WHITE}}$$

相反,如果非洲裔美国人为参照组,则白人为虚拟变量,此时,二元回归系数 $B'$ 可表示为:

$$B'_{\text{WHITE}} = \bar{Y}_{\text{WHITE}} - \bar{Y}_{\text{BLACK}}$$

不论哪类被选为参照组，平均收入的差异绝对值是不变的。

为一个多分类变量选择参照组(例如职业)多少有些复杂。对于职业这一变量，其所有的虚拟变量的回归系数的估计都是相对于该参照组的。尽管我们可以使用任意一个组作为参照组进行回归估计，从而产生与其他类的比较，但是仍有些准则需要了解，这些准则对解释回归估计非常有用：

第一，参照组应该明确界定。用其他剩余类可能不是一个好的选择，因为该选择无法确定“剩余”类的组成是什么。而且，我们感兴趣的组差异可能无法通过这些更相似的类与该剩余类的比较而反映出来。因此，选择一个明确的参照点，可以将组差异清楚地表现在方程中，这点非常重要。

第二，当定性类别中存在隐含次序关系时(例如职业)，有些研究人员通常选择最低或者最高级别作为参照组，而其他人更倾向于选择中间的类。尽管前者可以提供一系列系数估计来解释那些相对最低或最高级别的类，但是后者却可以减少那些粗心的研究者光抓住一个统计上显著的系数(例如职业)而忽略了其作为一个多类别的预测因子，在整体上是否有着显著影响的可能性(该问题会在第 4 章结尾详细讨论)。

第三，一个参照组要包含足够多的事件及信息，从而可以更加合理、准确地估计组群的均值。有时，研究者为了使其他类的信息保持“纯”，可能会选择只含少量观察值的类作为参照组。但是我们要明确，这个战略只可以用在参照组是高密度分布的类上，即在该变量所有的类中，其观测数最多或者为最多之一。

读者需谨记在心的是，在统计层面上，参照组是可以任意选取的。假设参照组是研究者根据适当的解释和推理的过程选出来的，那么就不会有“错误”的选择。而在实际层面上，“最佳”的选择是最大限度地减少额外计算的数量，因为这些额外计算会产生最具实质性利益的信息。

表 2.1 提供了如上所述的编码过程。对于种族这个变量，我们选择“白人”作为参照组；对于职业，我们选择“高级白领”作为参照组。受访者为非洲裔美国人的被编码为 1，白人受访者被编码为 0。$OCC_2$、$OCC_3$、$OCC_4$、$OCC_5$ 和 $OCC_6$ 是 5 个虚拟变量，它们旨在捕捉变量职业的 6 类里的全部信息。其中，$OCC_2$ 代表初级白领，$OCC_3$ 代表技术工人，$OCC_4$ 代表操作工人，$OCC_5$ 代表服务

业工作者，$OCC_6$为劳工。对于职业，事件 3、事件 14 和事件 15 均为 0，因为这些事件的受访者都是高级白领。而我们又可看出，事件 3 和事件 14 所有变量的编码在表中均为 0，原因在于，这两个事件的受访者均是种族和职业这两个定性变量的参照组中的成员，即他们都是从事高级白领职业的白人。

**表 2.1 种族及职业的虚拟变量的编码**

| 事件 | 种族 | 职业 | 黑人 | $OCC_2$ | $OCC_3$ | $OCC_4$ | $OCC_5$ | $OCC_6$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 黑人 | 初级白领 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 白人 | 技术工人 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 | 白人 | 高级白领 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 黑人 | 操作工人 | 0 | 0 | 0 | 1 | 0 | 0 |
| 5 | 黑人 | 劳工 | 0 | 0 | 0 | 0 | 0 | 1 |
| 6 | 白人 | 初级白领 | 0 | 1 | 0 | 0 | 0 | 0 |
| 7 | 白人 | 技术工人 | 0 | 0 | 1 | 0 | 0 | 0 |
| 8 | 白人 | 服务业工作者 | 0 | 0 | 0 | 0 | 1 | 0 |
| 9 | 黑人 | 服务业工作者 | 1 | 0 | 0 | 0 | 1 | 0 |
| 10 | 白人 | 初级白领 | 0 | 1 | 0 | 0 | 0 | 0 |
| 11 | 白人 | 操作工人 | 0 | 0 | 0 | 1 | 0 | 0 |
| 12 | 白人 | 初级白领 | 0 | 1 | 0 | 0 | 0 | 0 |
| 13 | 黑人 | 技术工人 | 1 | 0 | 1 | 0 | 0 | 0 |
| 14 | 白人 | 高级白领 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 黑人 | 高级白领 | 1 | 0 | 0 | 0 | 0 | 0 |

所有隐含在种族和职业中的定性信息都可以被转化成可供计算的信息。通过转化，我们可以计算集中趋势、分散程度、相关度及回归系数。我用 $j-1$ 个虚拟变量而不是原先含有 $j$ 类的分类变量，其中一个重要的原因在于，每一个虚拟变量可以从原先的度量中提取一部分信息。例如，每个虚拟变量记录着一个职业特征的存在与否（例如，1 代表其劳工的职业特征存在，0 则代表其劳工的职业特征不存在）。我们没有从根本上改变包含在种族或者职业的信息中的内容，我们只是选择了一个可替换的形式来表述这种信息。因此，只要我们可以调整对回归系数的诠释，使它们与自变量所隐含的测量性质相一致，那么就可以说，我们的统计基础非常坚实而且牢固。

# 描述性统计

## 分布统计

由于虚拟变量通常与定性度量相关,那么那些关于各类别的频数及其所占比例的描述性信息就对描述变量分布非常有用。同样,集中趋势的两个最常用的度量——众数和平均值,也可以提供很多有用的信息。

虚拟变量均值可以告诉我们被编码为 1 的类占所有分类的比例。回想一下,这个比例其实是一个相对频数,它是通过给定的分类事件数除以所有事件数($n_i/N$)得来的。让我们再回想一下那个计算均值的公式,即所有该度量下事件的数值的和除以事件的总数。假设所有的事件对一个虚拟变量只被编码成 0 或 1,那么加和所有事件的数值与统计所有被编码为 1 的事件数是等价的。因此,对于虚拟变量,比例公式与均值公式是等价的。

同样,对于包含连续度量的虚拟变量,其方差方程则与我们普遍运用的方差方程有关。

$$(\sum X_i^2)/N-(\sum X_i/N)^2 = n_j/N-p_j^2 = p_j-p_j^2 = p_j(1-p_j) \quad [2.1]$$

若 $X_i$ 是连续的,方差公式为第一个方程等号左边的部分。当把同样的方程运用到虚拟变量中时,$(\sum X_i^2)$ 变为 $n_j$,即被编码为 1 的事件数。第二项均值的平方 $(\sum X_i/N)^2$ 变为被编码为 1 的事件的比例,如上所述,这两者也应该是等价的。因此,我们可以证明,虚拟变量的方差其实是被编码为 1 的事件的比例与被编码为 0 的事件的比例的乘积。

当事件均匀地在两类之间分布时,虚拟变量的变异最大。现在,让我们来讨论些有关公众舆论的问题。“你是否支持为公共教育增加税收?”对于这个问题,当人们的意见均匀分布的时候,对税收政策持反对意见的最多。这时,选中任意两个都为“支持”的概率是最小的。当观点趋同时,即同意或者反对的概率接近 100%时,意见的多样性(或者说变异)会随之下降。

## 相关性

研究者不仅对那些可以描述单变量分布的度量感兴趣，还对变量之间的相联度量感兴趣。尤其当要用列联表分析来调查定性变量之间的关系时，我们会通过检验离散变量分类的均值差异来估测定性和定量变量之间的关系。最初，我们会将其限制到 3 个度量里，因此，我们可以看到按职业和种族分类后的平均收入差异(见表 2.2)。

**表 2.2　不同种族、职业下的收入平均值和标准差**

| | 平均收入 | 黑人的百分比 |
|---|---|---|
| **种族** | | |
| 白人<br>($N=2290$) | 7821.90<br>(4974.8) | |
| 黑人<br>($N=921$) | 4619.00<br>(2428.1) | |
| **职业** | | |
| 高级白领($OCC_1$)<br>($N=644$) | 10702.10<br>(7166.5) | 6.8 |
| 　白人<br>　($N=602$) | 10960.30<br>(7273.2) | |
| 　黑人<br>　($N=42$) | 7001.80<br>(3874.5) | |
| 初级白领($OCC_2$)<br>($N=337$) | 7680.90<br>(4228.7) | 17.1 |
| 　白人<br>　($N=279$) | 8061.30<br>(4462.6) | |
| 　黑人<br>　($N=58$) | 5850.80<br>(2039.9) | |
| 技术工人($OCC_3$)<br>($N=810$) | 6945.00<br>(2864.9) | 17.7 |
| 　白人<br>　($N=665$) | 7334.70<br>(2786.9) | |
| 　黑人<br>　($N=145$) | 5157.80<br>(2526.0) | |
| 操作工人($OCC_4$)<br>($N=788$) | 5553.90<br>(2454.1) | 38.9 |
| 　白人<br>　($N=481$) | 6085.30<br>(2414.6) | |

续表

| | 平均收入 | 黑人的百分比 |
|---|---|---|
| 黑人<br>($N=307$) | 4721.40<br>(2281.5) | |
| 服务业工作者($OCC_5$)<br>($N=287$) | 4434.4<br>(2352.0) | 51.2 |
| 白人<br>($N=139$) | 4805.6<br>(2626.5) | |
| 黑人<br>($N=148$) | 4085.8<br>(2008.3) | |
| 劳工($OCC_6$)<br>($N=345$) | 4090.0<br>(2020.1) | 64.0 |
| 白人<br>($N=124$) | 4777.30<br>(1900.1) | |
| 黑人<br>($N=221$) | 3704.30<br>(1986.6) | |

通过对表 2.2 的描述性分析，我们会发现 3 个非常明显的趋势：第一，黑人的平均收入比白人低；第二，从高级白领到劳工，平均收入是逐渐减少的；第三，从高级白领到劳工，黑人的比例是逐渐增加的。现在，我们必须寻找合适的方法来总结这 3 个二元关系，检验它们的显著性，然后通过控制其他相关因素来估测这些关系。

我们已知虚拟变量的均值和方差均和 $P_j$ 有关，那么，建立在样本方差和协方差上的相关度量也和虚拟变量的比例分布有关。大家知道，相关系数常用来测量变量之间相联度量的参数，且建立在两个变量之间协方差上的度量又和两个变量分布的离散程度有关。因此，我们可以说，两个定性变量度量的相关性对原始分布中方差的数值很敏感，因为虚拟变量的方差是 $P_j$ 的函数，且所涉及的虚拟变量相关性的强度会反映各类别出现频数的相对大小。

表 2.3 包含了由种族、职业及因变量收入构建的虚拟变量的零阶相关性系数估测，最右列列出了每个虚拟变量与收入的相关性，我们可以从其第二列相关系数看出。第一个数值(−0.313)表示虚拟变量黑人和收入的关系。负号说明编码为 1 的黑人虚拟变量与低收入相关，即美国非洲裔男性的平均收入比其

**表 2.3 种族、职业和收入之间的相关性系数**

| | 初级白领 ($OCC_2$) | 技术工人 ($OCC_3$) | 操作工人 ($OCC_4$) | 服务业工作者 ($OCC_5$) | 劳工 ($OCC_6$) | 收入 |
|---|---|---|---|---|---|---|
| 黑人 | −0.087*** | −0.139*** | 0.131*** | 0.157*** | 0.272*** | −0.313*** |
| $OCC_2$ (初级白领) | | −0.199*** | −0.196*** | −0.108*** | −0.119*** | 0.057*** |
| $OCC_3$ (技术工人) | | | −0.328*** | −0.181*** | −0.199*** | 0.007 |
| $OCC_4$ (操作工人) | | | | −0.178*** | −0.197*** | −0.166*** |
| $OCC_5$ (服务业工作者) | | | | | −0.108*** | −0.170*** |
| $OCC_6$ (劳工) | | | | | | −0.211*** |
| 平均值 | 0.106 | 0.250 | 0.244 | 0.089 | 0.107 | 6890 |
| 标准差 | 0.308 | 0.433 | 0.430 | 0.285 | 0.309 | 4622 |

注：*** 表示相关性系数在 0.001 显著性水平上统计显著。

他男性的平均收入低。通过对相关系数加平方，我们可以算出收入的样本方差有多少是可以被种族解释的。在这种情况下，我们知道，大约有 10%收入的方差可以通过组平均收入差异被解释。

由于在职业分类中需要多于一个虚拟变量来捕捉职业的所有信息，因此我们用 5 个相关系数来描述职业与收入的关系，每个都是针对一个特定的职业类别。这 5 个零阶相关系数的任何一个都可以估计该指定类(例如，$OCC_5$ 中服务业工作者)同其他所有类合并的收入差异。例如，$OCC_5$ 和收入之间的相关性系数为−0.170，它代表服务业工作者比非服务业工作者($OCC_1 + OCC_2 + OCC_3 + OCC_4 + OCC_6$)的工资低。当我们把该系数平方后，就可以估计，有百分之多少的收入方差是由于从事某一职业的男性比不从事该职业的男性挣得多或少这一事实引起的。这里，我们可知，有 2.89%的收入方差是由于服务业工作者比非服务业工作者的收入少引起的。还需注意的是，$OCC_3$(技术工人与非技术工人)和收入之间的相关性很小且不显著。从该度量可看出，技术工人与

非技术工人的平均收入没有显著的不同，该发现与技术工人不论在职业还是收入分布上都处于中间水平的事实一致。

其余各列的相关性表示每两个虚拟变量之间的相关度。因为两个虚拟变量之间的相关度等价于 φ 系数，又 $\varphi=(\chi^2/N)^{1/2}$，所以它们也与 $\chi^2$ 有关。任意两个度量之间的关系都可以在一个 2×2 的表格里表示出来。我们看到表格第一行涵盖了黑人与任意职业类别的相关性信息。一方面，我们发现 $OCC_2$ 与黑人呈现了负相关，这说明黑人在初级白领里的比例比在其他职业类别中的小；另一方面，黑人与 $OCC_4$、$OCC_5$ 和 $OCC_6$ 正相关，这说明黑人在操作工人中的比例(38.9%)、在服务性工作者中的比例(57.2%)和在劳工中的比例(64.0%)均比黑人不在操作工人中的比例(25.1%)、不在服务性工作者中的比例(26.3%)和不在劳工中的比例(24.3%)高。我们还可以从表格里看出，黑人与 $OCC_6$(劳工)的相关性最强，其原因在于黑人在劳工中的比例是最大的。[2]

## 偏相关

在其他自变量被控制时，我们可以通过偏相关系数估计一个因变量和一个自变量之间的关系。表 2.4 记录了当不断有自变量被控制时，$OCC_2$(初级白领)和收入($Y$)之间的一系列偏相关系数。

**表 2.4 职业虚拟变量和收入的偏相关系数及半偏相关系数**

| | | | |
|---|---|---|---|
| $r_{y,\text{occ2}}$ | =0.057*** | | |
| $r_{y,\text{occ2.occ3}}$ | =0.060*** | | |
| $r_{y,\text{occ2.occ3,occ4}}$ | =0.011 | | |
| $r_{y,\text{occ2.occ3,occ4,occ5}}$ | =−0.068*** | | |
| $r_{y,\text{occ2.occ3,occ4,occ5,occ6}}$ | =−0.171*** | | |
| | 偏相关系数 | 半偏相关系数 | 半偏相关系数的平方 |
| $r_{y,\text{occ2.occ3,occ4,occ5,occ6}}$ | −0.171*** | −0.191*** | 0.036 |
| $r_{y,\text{occ3.occ2,occ4,occ5,occ6}}$ | −0.271*** | −0.294*** | 0.087 |
| $r_{y,\text{occ4.occ2,occ3,occ5,occ6}}$ | −0.369*** | −0.387*** | 0.150 |
| $r_{y,\text{occ5.occ2,occ3,occ4,occ6}}$ | −0.337*** | −0.357*** | 0.127 |
| $r_{y,\text{occ6.occ2,occ3,occ4,occ5}}$ | −0.378*** | −0.394*** | 0.155 |

注：*** 表示相关性系数在 0.001 显著性水平上统计显著。

鉴于 $OCC_2$ 和收入之间的零阶关系是根据初级白领与其他类的工作者的平均收入相比较而来的，那么一阶偏相关系数 $r_{y,\ occ2.\ occ3}$ 就是控制了技工这一类得来的。由于技工现在被控制，那么这个偏相关代表初级白领且为除去初级白领和技工这两类工作者的平均收入差异后得来的。由于下一个系数控制了两个自变量（$OCC_3$ 和 $OCC_4$，即技术工人和操作工人），因此是一个二阶系数。在该例中，偏相关系数并不显著，这说明了初级白领的平均收入与除去了其本身以及技术工人和操作工人后的工作者（高级白领、服务工作者和劳工）的平均收入没有显著差别，这个结果的产生无疑是由于高收入组与低收入组是通过它们的中距均值求得的。随着越来越多的职业类别被控制，偏相关系数的阶数越来越高，数值为负的程度越大。最高阶数（或称“四阶偏相关”）控制了所有的职业虚拟变量，结果显示，初级白领的平均收入明显低于高级白领（参照组）的平均工资，此时，高级白领是唯一未被控制的组。表 2.4 的下半部分为我们呈现了职业虚拟变量的所有四阶偏相关系数。对于每一行，偏相关系数都表示一对职业虚拟变量和收入之间关系，该相关系数是通过消除其他变量的影响，比较高级白领（参照组）和指定职业类得来的。从纵向看，越往下，偏相关系数负的程度越大，这是因为当劳工和高级白领相比时，其收入差异的强度是最大的，而当初级白领与高级白领相比时，其收入差异是最小的。

表 2.4 的中间列和右边列记录了半偏系数和半偏系数的平方。我们知道半偏相关系数是建立在相关性和回归之间的有益桥梁。用于建立偏相关系数的、不断消除其他变量影响的剔除过程与用于建立偏回归系数的过程一样，会影响因变量和自变量的分布。然而，如果用半偏，本身对因变量有一定影响的自变量就不会因为其被控制而将这部分的影响剔除（Cohen & Cohen，1983）。半偏系数的平方表示一个自变量对 $Y_i$ 的可解释方差的唯一贡献。在这里，“唯一贡献”是指，$Y_i$ 的方差只归因于一个自变量，而不与其他被控制的自变量分享。例如，表示收入与 $OCC_2$ 的第一个半偏系数的平方数值 0.036 是在控制了 $OCC_3$、$OCC_4$、$OCC_5$ 和 $OCC_6$ 之后得出的。通过定义初级白领是

与高级白领不同的职业类别后，我们可以解释 3.6％的收入方差。换句话说，3.6％的收入方差可被解释是基于初级白领的平均工资比高级白领低这一事实。同样，15.5％的收入方差可以通过指定劳工与高级白领之间的收入差异来解释。因此，我们可知，保持其他条件一致，组间差异越大，所得到的可解释方差就越大。[3]

# 第 3 章 | 虚拟变量回归

在本章中,我们会探究 4 个含有虚拟变量的回归模型。最简单的模型可表述为因变量收入作为一个虚拟变量的线性函数。第二个模型和第一个相似,表述了收入作为一个单解释性特征的函数,但是在该模型中,自变量是职业,而在第一个模型中,自变量是种族。然而,由于职业是多级的,即其中含有两个以上的类,因此,我们需要构造 5 个虚拟变量。在第三个模型里,两个定性变量均会加入,因此我们可以讨论种族之间的收入差异是否可在与种族有关的职业差异里体现。最后一个模型,我们进一步扩展了自变量的数目,即定量的解释变量,连同职业和种族的虚拟变量都被包含在其中。

通过二元回归方程估计,我们可以确定对于给定数值的自变量,因变量的期望值是否会有所不同。因此,回归方程将均值从一个代表因变量期望值 $E(Y_i)$ 的单一点拓展到由一系列连续数值组成的线上。线上的每一个点都估计了一个特定的 $X_{ki}$ 条件下的 $Y_i$ 的期望值,表示为 $E(Y_i \mid X_{ki})$。这个系列的期望值是连续的,原因在于,$X_{ki}$ 本身是一个连续度量,其代表了无限的潜在数值。

当我们处理虚拟变量时,自变量为只有两个可能的数值的离散度量。那么,在建模时,尤其是为含有一个虚拟变量的连续因变量构造函数时,我们无法作出回归线。因此,我们可对每个可能的数值计算出一个 $Y_i$ 的期望值,即,当 $D_{ji} = 1$ 时,得到一个期望值;当 $D_{ji} = 0$ 时,再得到一个期望值。这些估测出来的数值相当于条件均值,即组群 $j$ 的均值 $Y_i$。

请看下面 3 个模型:

模型 1:$Y_i = f(\text{种族}) = \beta_0 + \beta_1 \text{BLACK} + u_i$

模型 2：$Y_i = f(\text{职业}) = \beta_0 + \beta_1 OCC_2 + \beta_2 OCC_3 + \beta_3 OCC_4 + \beta_4 OCC_5 + \beta_5 OCC_6 + u_i$

模型 3：$Y_i = f(\text{种族,职业}) = \beta_0 + \beta_1 BLACK + \beta_2 OCC_2 + \beta_3 OCC_3 + \beta_4 OCC_4 + \beta_5 OCC_5 + \beta_6 OCC_6 + u_i$

## 对含有一个虚拟变量的模型进行线性回归

在模型 1 里，通过收入对虚拟变量黑人进行回归，确定种族对于收入是否为一个有用的预测因子。表 3.1 列出了其回归结果。对这些连续和离散的自变量回归系数的合适解释要看常数项($B_0$)，其代表了当所有的自变量都为 0 时，$Y_i$的期望值；同时，还要看 $B_1$，它代表了每变化一个单元的 $X_k$，$Y_i$期望值的变化。当 $X_k$是连续的时候，$Y_i$的分布也是连续的，因此，其回归系数也可表示为斜率。相反，当 $X_k$为虚拟变量时，$Y_i$的变化是随每次 $B_k$单位的变化而变化，该变化与之前不同的是，其基于是否为指定类别的成员定义，因为虚拟变量一个单位的变化(从 0 到 1 或者从 1 到 0)反映了它是否属于某个指定类别的成员。

在此例中，虚拟变量黑人的回归系数是负值。这说明，黑人的预测收入比白人的预测收入少 3202.90 美元。$Y_i$的预测收入是通过简单的加减计算出来的。当 BLACK = 1 时，预测收入等于 $B_0 + B_1$，或者说 7821.9 − 3202.9 = 4619 美元；当 BLACK = 0 时，其预测收入就等于 $B_0$或者 7821.9 美元。读者可以发现这些预测值与表 2.2 中所列出的组群均值相等。

**表 3.1 模型 1、模型 2、模型 3 的回归结果**

| | 模型 1 | 模型 2 | 模型 3 |
|---|---|---|---|
| 常数 | 7821.9<br>(91.9) | 10702.1<br>(160.8) | 10811.4<br>(158.9) |
| 黑人 | −3202.9<br>(171.6) | | −1676.0<br>(172.4) |

续表

| | 模型 1 | 模型 2 | 模型 3 |
|---|---|---|---|
| $OCC_2$ | | −3021.2<br>(274.4) | −2842.1<br>(271.1) |
| $OCC_3$ | | −3757.1<br>(215.5) | −3566.4<br>(213.3) |
| $OCC_4$ | | −5148.2<br>(216.8) | −4604.5<br>(220.9) |
| $OCC_5$ | | −6267.7<br>(289.7) | −5512.7<br>(295.9) |
| $OCC_6$ | | −6612.1<br>(272.3) | −5647.8<br>(286.2) |
| $R^2$ | 0.09792 | 0.22400 | 0.24624 |
| $F$ | 348.3 | 185.0 | 174.1 |
| $R^2$ 的增量($R_3^2-R_1^2$) | | | 0.148 |
| 变化过的 $F$ | | | 126.1 |

模型 2 回归系数的方差、协方差矩阵

| | $OCC_2$ | $OCC_3$ | $OCC_4$ | $OCC_5$ | $OCC_6$ |
|---|---|---|---|---|---|
| $OCC_2$ | 75309.07 | | | | |
| $OCC_3$ | 25870.70 | 46439.50 | | | |
| $OCC_4$ | 25870.70 | 25870.70 | 47013.76 | | |
| $OCC_5$ | 25870.70 | 25870.70 | 25870.70 | 83922.03 | |
| $OCC_6$ | 25870.70 | 25870.70 | 25870.70 | 25870.70 | 74162.67 |

注：括号里为所估计的回归系数的标准误。

虚拟变量的显著性检验要遵循标准化过程。黑人的回归系数测量了黑人的期望收入相对于白人带来的影响如何。因此，黑人回归系数的标准误提供了白人与黑人的期望收入差异的标准误。当检验零效应这一零假设时，即由组群差异引起的期望收入差异不存在时，$t$ 检验所得数值可减少为相对标准误的回归系数的比率。同样，由于模型 1 包含了一个单一自变量，那么 $F$ 检验在此即对零假设的检验，其值为 $t$ 值的平方。$R^2$ 说明，种族差异解释了大约 10%的收入中的方差，这一点在之前的零阶相关系数检验中就已得知。

该例描述了当自变量为虚拟变量时，其回归结果解释和自变量为定量变量的相似与不同。常数项估计了参照组(白人)的期望收入；$B_1$ 估计了虚拟变量显

示出来的特征对期望值的影响(例如,黑人对其期望收入的影响),该影响捕捉了黑人与白人之间的收入差异。因此,零假设($\beta_1 = 0$)可表述为:$H_0$:$\mu_{BLACKS} - \mu_{WHITES} = 0$。对该模型的 $B_1$ 的 $t$ 检验和对模型 1 的 $F$ 检验基本上是相同的。[4]

## 对含有多个虚拟变量的模型进行回归

模型 2 估计了收入作为职业的函数,其中职业由 5 个虚拟变量表示。回归结果显示在表 3.1 的中间列。与模型 1 的解释一样,常数项 10702.1 为高级白领(参照组)的期望收入。其他的回归系数估计了从事相应职业类相对于高级白领的差异程度,从 $OCC_2$ 的回归系数可以看出,初级白领收入平均比高级白领少 3021.20 美元,为 7680.90 美元。相比之下,劳工平均比高级白领少挣 6612.10 美元,只有 4090 美元。

用一系列虚拟变量比用单一虚拟变量更能捕捉各职业组之间的差别信息,那么相应的职业对收入影响的显著性检验应该为该模型的 $F$ 检验。模型 2 的零假设可以写成 $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$,$F$ 检验是检验所有职业的期望收入是否都是一样的。另外,因为 $F$ 检验可以表达为 $R^2$ 与 $k$ 个自由度的商和 $1 - R^2$ 与 $N-k-1$ 个自由度的商的比率,其中,$k$ 等于自变量的数目,因此,$F$ 检验还可以看成是对 $R^2$ 的显著性检验。因此,拒绝零假设说明一个非零的收入方差可被受访者的职业所解释。基于模型 2 的回归结果,我们有:

$$F_{53205} = \frac{0.22400/5}{(1-0.22400)/3205} = 185.0$$

该数值在 0.001 的显著水平下非常显著。[5] 建立了职业的统计显著性之后,我们现在可以转移到对单个回归系数的 $t$ 检验上,从该检验我们可以看到,每个职业类的期望收入都与参照组呈现出显著不同。

## 估计类别之间的差异

$t$ 检验同虚拟变量的回归系数一起,使我们可以检验相对于参照组,某职业

类别带来的影响的显著性如何。然而,我们无法立即知道所比较的类别之间是否不同。例如,$OCC_6$(劳工)的回归系数负的程度最大,因此其期望收入也就最低。但是,我们如何知道劳工的期望收入确实是比服务业工作者或者操作工人低呢?

因为 $\beta_j = E(Y_i \mid OCC_j = 1) - E(Y_i \mid ref)$,所以期望收入在某两类之间的差异等于它们的回归系数之差($\beta_j - \beta_k$),其中,$\beta_j$ 为代表第 $j$ 类的虚拟变量的回归系数,同样,$\beta_k$ 代表第 $k$ 类的虚拟变量的回归系数。为了检验 $OCC_4$ 和 $OCC_6$ 的差异,即比较作为操作工人和劳工带来影响的差异,或者说,相对于劳工,操作工人所带来的影响,我们就必须用一个 $t$ 检验来估计回归系数之间的差异:

$$t = (B_j - B_k)/[\text{var}(B_j) + \text{var}(B_k) - 2\text{cov}(B_j B_k)]^{1/2} \qquad [3.1]$$

因为回归系数方差正好是标准误的平方,所以它们本身就容易得到。另外,许多统计软件包都有计算回归系数的方差、协方差矩阵的选项,研究者可以更加轻松灵活地完成这些额外检验。[6]

将 $OCC_4$ 和 $OCC_6$ 的估计值代入方程 3.1,我们有:

$$\begin{aligned} t &= -6612.1 - (-5148.2)/[74162.7 + 47013.8 - 2(25870.7)]^{1/2} \\ &= -1463.9/263.5 = -5.56. \end{aligned} \qquad [3.2]$$

在常用显著性水平 $\alpha = 0.05$ 下,我们知道,$t$ 的临界值为 $\pm 1.96$,从而我们可以推断劳工带来的影响确实与操作工人带来的影响不同,也就是说,劳工和操作工人确实有不同级别的平均收入。

## 第二个定性度量的加入

当我们回头看表 2.2 中组群的均值时,可以发现,从高级白领到劳工,下降的不仅仅是平均收入,同时还有白人的比例。我们想知道,当控制了职业的收入差异时,种族差异是不是还会在收入中存在。要回答这个问题,就需要检验虚拟变量黑人的偏回归系数。表 3.1 最右列的模型 3 给了我们一个比较满意的

答复。从常数 10811.4 可以得出当所有自变量为 0 时的期望收入，或者说，该常数即白人高级白领的期望收入。黑人的回归系数－1676.0 表示，在我们考虑了同职业类别有关的收入方差和黑人在各职业类别中不是均匀分布的事实后，黑人的平均收入比白人少 1676 美元。尽管这个值比模型 1 中估计的小一些，但是该估计的收入差异在 0.001 显著性水平下非常显著。黑人的回归系数强度随职业由高到低逐渐下降，该现象说明黑人的平均收入比白人低的一个原因在于，黑人大体上都集中在低收入职业类别里。同样，如果控制了收入和职业分布中的种族差异，偏回归系数连同职业虚拟变量一起，可以估计出每个指定类别的成员对期望收入的影响大小。

为了决定控制了职业后的种族局部效应或者控制了种族后的职业局部效应是否仍在统计上显著，我们还要用 $F$ 检验来确定。与其单靠 $F$ 检验来检验整个方程，不如在控制了其他变量后，用增量 $F$ 检验来检验一个或一组分类变量的解释功效。例如，我们可以把模型 3 看成模型 1 和模型 2 的结合，我们在模型 3 里加入了代表职业类别中种族差异的虚拟变量。和之前一样，职业定位的解释功效是由一组虚拟变量而不是一个虚拟变量来捕捉的，因此，我们可以通过比较模型 1 和模型 3 的 $R^2$ 值，或者平均回归平方和来估计职业的分布。零假设在这里可表示为 $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$，换句话说，一旦我们控制了收入和各职业类别中的种族差异，期望收入在所有的职业类别中都是相等的。该 $F$ 检验的公式可表述为：

$$F = \frac{(R_3^2 - R_1^2)/(k_3 - k_1)}{(1 - R_3^2)/(N - k_3 - 1)} \qquad [3.3]$$

其中，$R_3^2$ 是模型 3 的 $R^2$ 值，$R_1^2$ 为模型 1 的 $R^2$ 值，$N$ 为事件数目，$k_1$ 和 $k_3$ 分别为模型 1 和模型 3 里自变量的数目。分子部分表明了相对于模型 1 和模型 3 自变量的数目差异，由职业类别影响所带来的 $R^2$ 增量。分母部分为当种族和职业都包括在内后，所剩的、不能被解释的方差的比例与相应的自由度的商是多少。在该例中，加上从 $OCC_2$ 到 $OCC_6$ 观测所得的 $R^2$ 增量为 0.14832，然后我们还需算出其除以 5 个自由度后所得的值。因此，在控制了种族后，用 $F$ 检验算

出职业类别的显著性为：

$$F = (0.14832/5)/(0.75376/3204) = 126.1$$

## 期望值

由于种族包含 2 类，职业包含 6 类，那么他们一起可以生成共 12 个不同的组群。模型 3 针对这 12 组群估计了他们的期望收入。这些估计值和 12 个组群的均值相等，即与各个种族在每个职业类别中的期望收入均值相等。通过用表 3.2 中描述的回归系数集合，读者可以自己计算出各组群的收入期望值，然后和表 2.2 中列出的数值进行比较。表 2.2 中列出各组群的期望值是根据模型 1 估计出的种族参数，或者是模型 2 估计出的职业类别参数所得出的。和之前的结果不同的是，从模型 3 算出来的 12 个组群的期望收入和表 2.2 中的数值非常不匹配，这是为什么呢？

这是由于在模型 1 和模型 2 中，我们把所有检验限制在一维中，即种族或者职业。当将观测拓展到模型 3 时，我们其实是基于一个简单假设，即"黑人对各职业类别的影响是一样的(例如，黑人和白人之间的期望收入有差异)"和"职业间的收入差异对黑人和白人也是一样的"。我们知道，当计算职业类别中黑人和白人的期望收入有差异时，这个假设已经开始奏效，即不论职业，黑人和白人工作者的差别总是 1676($B_1$ 的值)；不论种族，服务业工作者和高级白领之间的差异总是 −5512.7($B_5$ 的值)。这种等价性影响是模型阐述的结果。

**表 3.2　模型 3 的收入预测值**

| | Blacks | Whites |
|---|---|---|
| $OCC_1$ | $B_0 + B_1 = 9135.4$ | $B_0 = 10811.4$ |
| $OCC_2$ | $B_0 + B_1 + B_2 = 6293.3$ | $B_0 + B_2 = 7969.3$ |
| $OCC_3$ | $B_0 + B_1 + B_3 = 5569.0$ | $B_0 + B_3 = 7245.0$ |
| $OCC_4$ | $B_0 + B_1 + B_4 = 4530.9$ | $B_0 + B_4 = 6206.9$ |
| $OCC_5$ | $B_0 + B_1 + B_5 = 3622.7$ | $B_0 + B_5 = 5298.7$ |
| $OCC_6$ | $B_0 + B_1 + B_6 = 3487.6$ | $B_0 + B_6 = 5163.6$ |

当把简化过的假设加入到模型 3 的规范中后，经验上是否行得通呢？事实上，通过了解各组群的期望收入与表 2.2 中组群均值差异，我们可能想改变这个规范。我们会在模型 4 中提供一个比较正式的检验，其结果会告诉我们哪个模型更好，是模型 3 还是允许差异效应（例如，交互作用）的模型。

## 在规范中加入定量变量

用回归分析的好处是，即便一些自变量是分类变量，建模过程还是非常灵活的。基于这点，我们会将最初的模型限制在虚拟变量回归因子中，这样读者就可以习惯虚拟变量回归系数的解释方法了。在该部分，我们准备把定量和虚拟变量回归因子都加入观测中。因此，我们提出了模型 4——期望收入是种族、职业、教育和工作任期的函数。

$$
\begin{aligned}
\text{模型 4}：Y_i &= f(\text{种族，职业，教育，工作任期}) \\
&= \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 + \beta_4 \text{OCC}_4 \\
&\quad + \beta_5 \text{OCC}_5 + \beta_6 \text{OCC}_6 + \beta_7 \text{EDUC} + \beta_8 \text{TENURE} + u_i
\end{aligned}
$$

EDUC 和 TENURE 都是用年来衡量的定量变量。由表 3.3 可见模型 4 的回归结果。

我们可以看出，该模型估计结果的常数比之前的估计都小。更重要的是，模型设定的变化也改变了其实质性的意义。现在的常数估计的是那些没受过教育且工作任期为 0 的白人高级白领的期望收入，这些特征几乎是不可能存在的。虚拟变量黑人的回归系数现在表示当把职业、教育和工作任期这些对收入方差有影响的变量剔除后，黑人和白人之间的期望收入差异（1188.10 美元）。虚拟变量职业的回归系数估计了在控制了其他自变量之后，每个职业类别相对于参照组的期望收入的净差异。比如，初级白领平均比高级白领的收入少 2316.10 美元等等。同样，当保持种族、职业和教育不变时，每增加一年的工作任期，收入可增加 84.70 美元。对于教育，每增加一年的教育经历，期望收入会增加 282 美元。

由于另外两个定量变量的加入，我们可以把模型 4 想象为，其生成了一系列

的回归平面，并引入了截距、斜率和偏斜率的概念。在文中，我们还可以将虚拟变量的回归系数表示成不同的截距。由于斜率或者偏斜率只会与定量自变量相关，因此，与 EDUC 和 TENURE 有关的回归系数提供了对偏斜率的估计。图 3.1 描述了模型 4 的回归结果。为了便于读者比较各组群之间教育和工作任期的截距、斜率和偏斜率，图 3.1 没有用三维空间图，而是用二维直线图对其进行了比较。又由于教育和工作任期均用年来衡量，则我们可以用它们相对于同一度量的局部效应而进行比较。

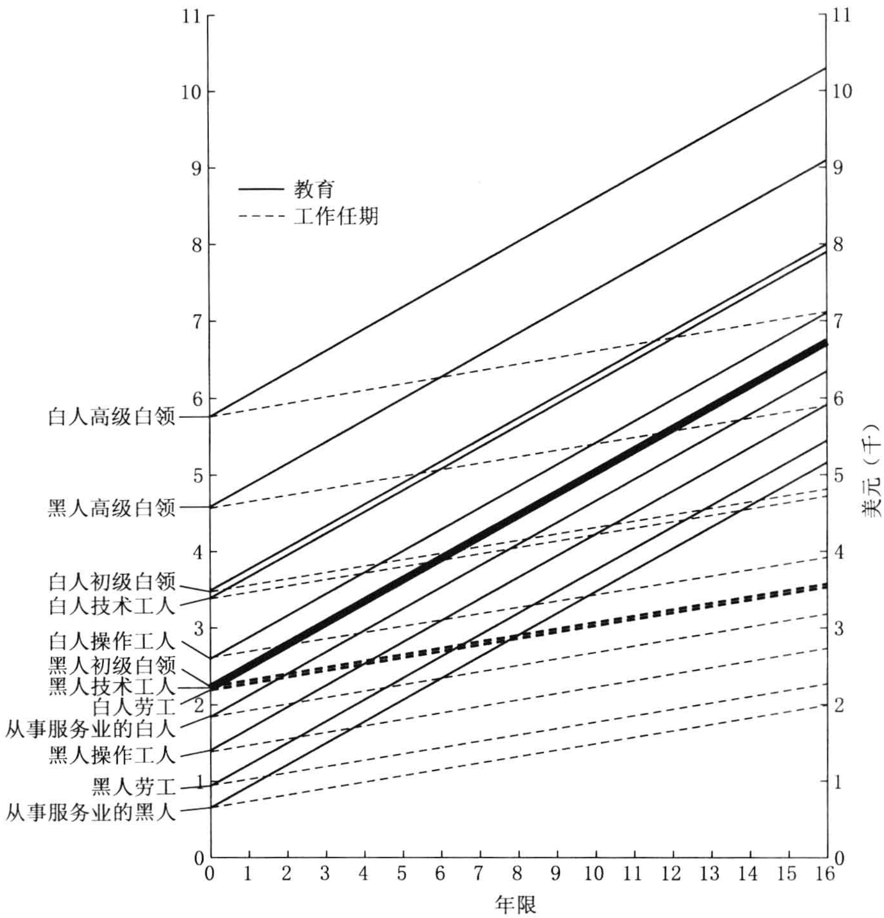



**图 3.1 模型 4 的回归结果**

对于所有受访者，由于模型设定把估计值限制到教育年限和工作任期所带来的平均影响里，因此所有实线均有着相同的斜率(282 美元/年)，所有的虚线也有着相同的相率(84.70 美元/年)。然而，Y 截距是可以因组群的不同而不同的，因此在图中有 12 个不同的截距。由表 3.2 可见，计算这些截距时，我们用了相同的回归系数组合。但是，因为回归系数本身随着模型设定拓展而改变，那么，计算出来的截距自然会与表 3.2 中列出的预测收入有所不同。从每个“按种族特征分类的职业组”(共 12 个)的截距中分出一个实线和一个虚线，其分别代表教育年限和工作任期的偏效应。对每个“按种族特征分类的职业类别”，我们就可以辨别出：(1)合适的 Y 截距，即有 0 年教育经历和 0 年工作任期的给定组群成员的 Y 的期望值；(2)对于一个特定的组群，每增加一年的教育经历，Y 净增加的期望值(实线)；(3)对于一个特定的组群，每增加一年的工作任期，Y 净增加的期望值(虚线)。就如模型 3，“各组群的等价性”仅是嵌入在模型设定中的一个假设，其经验性仍有待检验。

**表 3.3 模型 4 的回归结果**

| | | | |
|---|---|---|---|
| 常数 | 5761.1<br>(359.0) | $OCC_6$ | −3606.8<br>(306.4) |
| 黑人 | −1188.1<br>(169.4) | 教育 | 282.0<br>(23.1) |
| $OCC_2$ | −2316.1<br>(261.8) | 工作任期 | 84.7<br>(6.6) |
| $OCC_3$ | −2343.7<br>(223.7) | $R^2$ | 0.31459 |
| $OCC_4$ | −3166.6<br>(237.5) | $F$ | 183.7 |
| $OCC_5$ | −3918.5<br>(299.9) | $R^2$ 的增量(相对于模型 3)<br>变化过的 $F$ | 0.068<br>159.7*** |

注：括号里为标准误。

*** 表示相关性系数在 0.001 显著性水平上统计显著。

# 第 4 章 | 估计组影响差异

上一章的模型是拓展了模型设定中的自变量数目和种类后得到的。我们知道，所有含多变量的模型都有一个简化的假设，即任意自变量直接对因变量的影响与该自变量通过其他自变量对因变量的影响是一样的。换句话说，我们没有包含任何交互项来检验职业、教育或者工作任期是否在黑人和白人之间有所不同。在本章中，我们将建立一个新模型，通过引入交互项来检验假设的有效性，然后来回答之前讨论的两个问题——违反回归模型假设的后果和用非独立检验做多重比较的替代方法。

估计组群间的平均影响可以提供一个有用又简单的关系描述。然而有时，一个自变量($X_i$)通过第二个自变量($Z_i$)的分类或数值所产生的对因变量($Y_i$)的影响是不同的。当 $X_i$ 和 $Y_i$ 之间的关系由变量 $Z_i$ 决定时，我们就需要调整模型的规范，使 $X_i$ 和 $Y_i$ 之间的关系相对于 $Z_i$ 而改变。检验这种差异效应需要用交互项，即包括在模型设定里的两个或多个自变量的乘积。

交互项可以定义为两个定量变量的乘积，也可以是两个虚拟变量的乘积，或者是一个定量变量和一个虚拟变量的乘积。另外，更复杂的交互项可以包括两个以上的变量。根据文献，早期对该系列研究有卓越贡献的包括 Jaccard、Turrisi 和 Wan(1990)，当所有变量都是连续度量时，他们为所出现的交互作用提供了非常完美的解释。但是在本章，我们关注的是其他的交互组合。

可以考虑构造一个包含两个虚拟变量($D_{1i}$、$D_{2i}$)的交互项来衡量两个二分的定性变量，比如性别和婚姻状况。我们可能会假设，作为女性所带来的影响可能会由其是否结婚而决定。在此情况下，我们会检验作为女性所带来的影响大小、已婚所带来的影响大小和交互项，包含变量 $D_{1i}$、$D_{2i}$ 的乘积，即 $D_{1i}\times$

$D_{2i}$。这样，当一个受访者为已婚女性时，其交互项就等于1，交互项的回归系数估计了已婚女性和其他受访者的不同影响程度。其中，其他受访者包括了已婚男性、未婚女性和未婚男性。

现在，我们可以考虑另一个交互项，它由一个测量年龄的定量变量 $X_i$ 和一个性别虚拟变量 $D_{1i}$ 所定义。在这里，我们可能会假设年龄所带来的影响由性别决定。因此，我们会试图用模型来检验年龄所带来的影响($X_i$)、性别所带来的影响($D_{1i}$)以及交互项($X_i \times D_{1i}$)的影响。在这里，交互项对所有男性都为0；而对女性，则会默认为她们的年龄。该交互项的回归系数估计了年龄对女性的影响小于(或大于)年龄对男性的影响。

对于检验有关老年工作者收入数据的情况，图3.1可以帮我们把几种差异效应概念化。从图中我们可以看出，由于模型3和模型4的规范不同，使得黑人和白人之间的收入差异对所有职业类别都是一样的。那么，黑人的收入劣势会不会在高级白领中比较大，而在劳工中比较小呢？换句话说，$Y$ 截距之间的差异在高级白领中是不是比在劳工中更大呢？这个问题的一个更普遍的问法是："种族的影响是不是在各职业类别中不同？或者说，是不是从事不同职业对黑人和白人的影响是不同的？"

为了对那些包含定性变量的交互项作出更好的解释，我们可以先来看看表4.1，其展示了收入、种族和职业之间的3种可能关系。在3个部分里，均列出了不同种族在各职业类别里的平均收入，而且，每一列的边际值还列出了该职业类别的平均收入。我们发现，表格的前两个部分用的是可以说明交互作用类型的假设数据，还可以发现相同职业类别的均值都是相同的。知道了这些，我们就可以进一步发现表格最右端的黑人和白人之间的收入边际差异不总能准确地捕捉到每部分内种族、职业和收入之间的关系。

表格第一部分的一些数据是没有交互作用的。当交互作用不存在时，处于边际的黑人和白人的平均收入差异为3203美元，其与控制了职业类别后，黑人和白人在每一列内的平均收入差异的大小是一样的。读者可以通过用第二行每列的均值扣除第一行的均值来证实这点。在该情况下，所估计的种族在一个回归模型里的影响，即对每个职业组的平均影响(比如模型3)，就可以提供一个比较准确的

黑人、白人的收入差异,因为种族对所有职业类别的影响是一致的。

**表 4.1 交互项可能的种类**

| | 高级白领 | 初级白领 | 技术工人 | 操作工人 | 服务业工作者 | 劳　工 | |
|---|---|---|---|---|---|---|---|
| 无交互作用[a] | | | | | | | |
| 黑人 | 7708 | 5029 | 4315 | 3599 | 2883 | 2939 | 4412 |
| 白人 | 10911 | 8232 | 7518 | 6802 | 6086 | 6142 | 7615 |
| | 10702 | 7681 | 6945 | 5554 | 4434 | 4090 | |
| 包含强度和方向差异的交互作用[a] | | | | | | | |
| 黑人 | 7002 | 5851 | 5158 | 6111 | 5120 | 4628 | 5470 |
| 白人 | 10960 | 8061 | 7335 | 5198 | 3704 | 3130 | 7479 |
| | 10702 | 7681 | 6945 | 5554 | 4434 | 4090 | |
| 包含强度差异的交互作用[b] | | | | | | | |
| 黑人 | 7002 | 5851 | 5158 | 4721 | 4086 | 3704 | 4619 |
| 白人 | 10960 | 8061 | 7335 | 6085 | 4806 | 4777 | 7822 |
| | 10702 | 7681 | 6945 | 5551 | 4434 | 4090 | |

注:a. 构造的数据。
　b. 真实的数据。

表 4.1 的第二部分仍然是基于构造的数据,说明了种族影响在强度和方向上的不同。用这个交互类型,种族的影响不仅在不同职业类别中不同,即黑人和白人之间的平均收入差异强度不同,而且其影响方向也不同。假设样本中组群均值就如表格 4.1 中所列的一样,那么在本例中,种族的收入边际差异或者中距的平均影响为 2009 美元,从而可知,白人平均收入稍高。但是当我们再仔细观察表格的该部分,若以指定职业组间的比较差异为条件,那么,黑人和白人之间的收入差异的强度是不同的。这个差异在高级白领中最大(3.958 美元),在操作工人中最小(913 美元)。由于在各职业类别中,黑人和白人之间收入差异并不一致,因此我们需要指定一个交互项。同时,我们还会发现该差异不仅在强度上不同,其方向或者符号的正负也不总是相同。只有在高级白领、初级白领和技术工人中,白人的平均收入才比较高;相反,在操作工人、服务业工作者和劳工中,黑人的平均工资比较高。这即差异在方向上的转换,该转换有着特殊种类的交互作用的特征,因此黑人和白人收入的边际差异隐藏了一些种族差

异的重要部分。如果将各职业类别的影响平均化，那么在初级白领和技术工人中，其种族影响是合理且准确的。但是对于高级白领，平均化低估了种族差异，而且对于其他职业类别，平均化还错误地估计了差异的方向。

通过表 4.1 的第三部分，我们可以比较实际数据和构造数据的分布。这些收入均值也可以在表 2.2 中找到。在实际数据中，黑人和白人之间的收入边际差异为 3203 美元，我们已经在第一个回归模型里估计出该值。但是当我们把视线放到每个职业内时，会发现黑人和白人之间的收入差异是由职业类别决定的，该差异在高级白领中最大(3958 美元)，而在服务业工作者中最小(720 美元)。然而，收入差异总是在一个方向，即黑人工作者的平均收入总是比较小。在处理这种交互项时，其平均差异影响的方向通常是正确的，只是对于部分职业类别比较小，而在其他类别中比较大。

## 解释交互效应

为了检验交互效应，我们需要一个可以估计差异效应和确定其显著性的模型设定。要达到这个目标，我们就需要构造 5 个乘积项，然后把这 5 个项加入模型设定中。通过模型 5，我们可以检验不同种族在不同职业中的差异效应，或者说不同职业在不同种族中的差异效应。

模型 5：$Y_i = f$(种族，职业，教育，工作任期)

$$\begin{aligned} &= \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 + \beta_4 \text{OCC}_4 \\ &\quad + \beta_5 \text{OCC}_5 + \beta_6 \text{OCC}_6 + \beta_7 \text{EDUC} + \beta_8 \text{TENURE} \\ &\quad + \beta_9 \text{BLOCC}_2 + \beta_{10} \text{BLOCC}_3 + \beta_{11} \text{BLOCC}_4 \\ &\quad + \beta_{12} \text{BLOCC}_5 + \beta_{13} \text{BLOCC}_6 + u_i \end{aligned}$$

新变量 $\text{BLOCC}_2$ 到$\text{BLOCC}_6$ 是由虚拟变量黑人与每个职业虚拟变量相乘而得来的。如果受访者既是黑人又属于初级白领，那么$\text{BLOCC}_2$ 被编码为 1。因此，由$\text{BLOCC}_2$ 的回归系数估计出的平均收入的增加或者降低，只对这个组群，即黑人初级白领有用。

模型 5 的结果在表 4.2 里已列出。最初，我们可能想知道是否要在模型拟合中允许种族和职业的差异效应而使统计显著性提高。我们可以就这个问题，通过方程 3.3 的 $R^2$ 增量测试，比较模型 5 与模型 4 的结果：

$$F_{53197}=\frac{0.00679/5}{0.67862/3197}=6.4$$

**表 4.2 模型 5 的回归结果**

| | | | |
|---|---|---|---|
| 常数 | 5794.8<br>(358.7) | TENURE | 84.0<br>(6.6) |
| BLACK | −3793.3<br>(610.1) | $BLOCC_2$ | 1501.2<br>(823.0) |
| $OCC_2$ | −2274.9<br>(280.2) | $BLOCC_3$ | 2326.2<br>(705.0) |
| $OCC_3$ | −2418.4<br>(232.7) | $BLOCC_4$ | 2984.8<br>(672.5) |
| $OCC_4$ | −3427.2<br>(256.3) | $BLOCC_5$ | 3528.0<br>(761.0) |
| $OCC_5$ | −4513.4<br>(372.5) | $BLOCC_6$ | 3383.9<br>(747.3) |
| $OCC_6$ | −4202.8<br>(399.0) | $R^2$<br>$F$ | 0.32138<br>116.46 |
| EDUC | 292.9<br>(23.1) | $R^2$ 的增量($R_3^2-R_1^2$)<br>变化过的 $F$ | 0.007<br>6.42*** |

注：括号里为回归结果的标准误。
*** 表示回归系数在 0.001 显著水平上显著。

该 $F$ 值在 0.001 显著水平上统计显著。尽管所增加的解释功效没有显著到可以拒绝的地步，但是 $F$ 检验的确告诉我们，大样本量可使我们估计的差异效应更加合理准确。

现在我们再看对回归系数的解释。模型 5 的常数项与模型 4 中的一样，估计了有 0 年教育经历和工作任期的白人高级白领的预测收入。另外，EDUC 的回归系数告诉我们，在控制了工作任期、种族、职业以及职业内不同的种族影响后，教育对收入的平均影响是多少。TENURE 回归系数的解释与之类似，在此就不详细说明了。

黑人和职业虚拟变量的回归系数看上去像模型 4 的延续，其实不然。由于

引入了种族×职业的乘积项,它们的意义有了变化。当我们继续专注于由不同种族在不同职业中形成的12个组群时,可以通过将各回归系数映射到其所在的组群中,以弄清楚每个特定的回归系数所扮演的角色是什么。该映射结果已在表4.3中列出,其根据当组群中的个体在虚拟变量中(包括交互项里)被编码为1时,回归系数被加入到特定组群的估计中而得来。为简化起见,我们还是将EDUC和TENURE的回归系数设定为0。

**表4.3　不同种族预测收入系数**

| | 白　人 | 黑　人 |
|---|---|---|
| 高级白领 | $B_0$ | $B_0+B_1$ |
| 初级白领 | $B_0+B_2$ | $B_0+B_1+B_2+B_9$ |
| 技术工人 | $B_0+B_3$ | $B_0+B_1+B_3+B_{10}$ |
| 操作工人 | $B_0+B_4$ | $B_0+B_1+B_4+B_{11}$ |
| 服务业工作者 | $B_0+B_5$ | $B_0+B_1+B_5+B_{12}$ |
| 劳工 | $B_0+B_6$ | $B_0+B_1+B_6+B_{13}$ |

从黑人的回归系数开始,$B_1$估计了黑人高级白领和白人高级白领之间的期望收入差异。与模型4不同的是,其提供的不再是黑人在所有职业类别中的平均影响。系数的$t$检验是对零假设的检验,即在控制了由教育、工作任期带来的方差后,黑人高级白领与白人高级白领的期望收入是一样的零假设。换句话说,即在高级白领中,黑人对期望收入并没有显著影响。由于该回归系数的$T$值为−6.22,因此零假设可以被拒绝。然后我们知道,在高级白领中,当控制了模型中的其他因素后,黑人平均收入显著地低于白人。

同样,职业虚拟变量的回归系数也不再提供黑人和白人在某一特定职业类别相对于参照组的平均影响的估计。我们用$B_2$取而代之,即$OCC_2$的回归系数来估计白人初级白领与白人高级白领之间的期望收入差异,从回归中,我们知道,白人初级白领平均比白人高级白领的收入低2274.90美元。同样,白人操作工人平均比白人高级白领的收入低3427.20美元。换句话说,一旦指定了乘积项,原来变量的系数,即黑人在$OCC_2$到$OCC_6$的系数,成为包括了参照组影响的比较。比如,$B_1$测量的是职业类别参照组——黑人高级白领所带来的影

响；$B_2$ 到 $B_6$ 测量的是种族的参照组——白人在除了高级白领的某一特定职业类别中所带来的影响。$t$ 检验连同 $OCC_2$ 到 $OCC_6$ 的回归系数一起，测量的是白人工作者在各职业类别的收入差异显著性。表 4.2 的结果表明，对于白人，高级白领与其他职业类别的估计差异显著不同。

乘积变量的回归系数估计了不同种族从事不同工作的差异效应。同样，我们还可以通过这些回归系数来估计黑人在各职业类别里的差异效应。为什么这两种说法都可以接受呢？通过观察表 4.3，我们可以给出问题的答案。白人受访者从事初级白领和高级白领工作的预测收入差异可通过 $B_2$（$-2274.9$）来捕捉，对于黑人，这个差异可通过 $B_2+B_9$（$-2274.9+1501.2$）来捕捉。因此，$B_9$ 估计的是初级白领（相对于高级白领）在黑人与白人中的收入差异影响。因为 $BLOCC_2$ 的系数为正，所以黑人在初级白领与高级白领中的收入差异比白人少 1501.20 美元，或者说在黑人中，该收入差异是 $-773.70$美元而不是$-2274.90$美元。同样，黑人高级白领与白人高级白领之间的期望收入差异为 $B_1$（$-3793.3$），而黑人与白人在初级白领工作的期望收入差异为 $B_1+B_9$（$-3793.3+1501.2$），$B_9$ 估计了黑人相对于白人在初级白领与高级白领间的收入差异影响，$-3793.3$ 为黑人与白人在高级白领上收入的差异，而$-2292.1$ 为黑人与白人在初级白领收入上的差异。因此，黑人各职业的期望收入差异需要由两个回归系数捕捉，即 $\beta_j+\beta_{jk}$，其中 $\beta_j$ 为职业虚拟变量（$OCC_2$ 到 $OCC_6$），$\beta_{jk}$ 为乘积变量的系数。我们可以把 $\beta_j$ 和 $\beta_{jk}$ 之间的关系定义为如下：

$$\beta_j = E(Y_i \mid \text{WHITE, OCC}_j) - E(Y_i \mid \text{WHITE, OCCref}) \quad [4.1]$$

$$\beta_{jk} = [E(Y_i \mid \text{BLACK, OCC}_j) - E(Y_i \mid \text{BLACK, OCCref})] - [E(Y_i \mid \text{WHITE, OCC}_j) - E(Y_i \mid \text{WHITE, OCCref})] \quad [4.2]$$

$$= [E(Y_i \mid \text{BLACK, OCC}_j) - E(Y_i \mid \text{BLACK, OCCref})] - \beta_j \quad [4.3]$$

因此，

$$\beta_j + \beta_{jk} = E(Y_i \mid \text{BLACK, OCC}_j) - E(Y_i \mid \text{BLACK, OCCref}) \quad [4.4]$$

就像我们从方程 4.2 中看到的一样，乘积项回归系数的 $t$ 检验不是黑人在不同

职业中的期望收入净差异。我们所检验的假设要验证的是，对于黑人和白人，在指定职业类别与参照组间的收入净差异是否一样。

如果这些乘积项系数为负，我们就有证据说明，相对于白人，黑人高级白领与其他职业类别的收入差异更大。如果再加上 $OCC_2$ 到 $OCC_6$ 的负系数，那么黑人在各职业类别中的收入差异与白人相比就更加显著。但是，从回归结果看出，乘积项的系数是正值，那么在不同职业类别间，白人的收入差异相对而言就更显著了，因此，相比之下，黑人就没那么显著，或者说其各职业类别间的收入就更接近。事实上，黑人在不同职业间的收入可能没什么区别，而且当职业等级逐渐降低时，收入差异不论在黑人之间还是白人之间似乎都是缩小的。

现在，我们进一步检验一下这些结论。为了陈述之前的两个结论，我们必须确定当控制了教育和工作任期之后，非洲裔美国人在不同职业类别中的收入差异是否显著。通过计算机程序执行的回归系数 $t$ 检验，没有一个可以解决这个问题。交互项的 $t$ 检验值只会告诉我们，从事不同职业的净效对黑人和白人是否有显著不同。然而，知道作为一个技术工人、一个操作工人、一个服务业工作者或劳工，并不能确定从事某一职业类别的净效应，其作为一个可靠的预测因子一定比另一个职业类别好。要回答这个问题，我们则需再深入一点。

**表 4.4　不同种族所得收入的职业净效应**

| | 白　人 | 黑　人 |
|---|---|---|
| 初级白领 | −2274.9<br>(280.2) | −773.7<br>(776.7) |
| 技术工人 | −2418.4<br>(232.7) | −92.2<br>(683.5) |
| 操作工人 | −3427.2<br>(256.3) | −442.4<br>(646.5) |
| 服务业工作者 | −4513.4<br>(372.5) | −985.4<br>(682.6) |
| 劳工 | −4202.8<br>(399.0) | −818.9<br>(667.6) |

表 4.4 通过控制教育和工作任期,检验了黑人和白人的职业取向效应。这里,同白人有关的项与表 4.2 中 $OCC_2$、$OCC_3$、$OCC_4$、$OCC_5$、$OCC_6$ 的系数一样,这些系数估计的均是白人各职业类别与参照组——白人高级白领间的收入差异。检验这些系数的统计显著性是为了了解在所有人群中,白人工作者在各职业类别中的收入差异是否存在。另外,如果我们要检验白人操作工人平均收入是否比白人技术工人低,那么,我们就应控制种族、教育和工作任期,通过方程 3.1,模型 5 中的 $B_3=-2418.1$、$B_4=-3427.2$ 及它们各自的方差、协方差,来比较这两个职业类别的收入差异。

对于黑人,一个职业类别的效应由两个回归系数捕捉:职业虚拟变量的回归系数和交互项的系数。例如,当要确定黑人初级白领和高级白领在期望收入上如何不同时,我们会将白人高级白领和白人初级白领的期望收入差异 $B_2$ 和黑人相对于白人在高级白领和初级白领中的收入差异 $B_9$ 加起来比较。根据表 4.3,我们可以看出,要计算黑人初级白领的期望收入,就要用计算黑人高级白领期望收入的两个回归系数和另外两个系数——$B_2+B_9$。[7] 因此,对于黑人,表 4.4 中一个职业类别中的估测收入效应可以通过对适当的回归系数求和来构造。

正如我们在表 4.4 中看到的一样,黑人相对于白人,在非高级白领职业中显示的收入劣势要小一些。与白人的职业间收入差异不同的是,存在于黑人中最小的差异是高级白领与技术工人之间的差异,而不是高级白领与初级白领之间的差异。事实上,黑人高级白领与初级白领的收入净差异和高级白领与劳工的差异差不多。然而,这些关于黑人职业间的期望收入净差异仅基于回归系数。那么,回归检验可以使我们对这些估测的差异更自信吗?如果不可以,那么这些差异是否大多归因于抽样误差?

表 4.2 中乘积项系数的 $t$ 检验告诉我们,除了初级白领,任意指定职业类别和高级白领在黑人或白人中的收入差异都是显著不同的。但是,我们还没有直接检验黑人职业间收入差异是否显著。那么,必要的 $t$ 检验是估计两个回归系数和相对于该统计样本分布的标准差的值,即 $(B_2+B_9)/SE(B_2+B_9)$。下面这个与方程 3.1 相似的方程正好可以达到此目的:

$$t = B_j + B_{jk} / [\text{var}(B_j) + \text{var}(B_{jk}) + 2\text{cov}(B_j, B_{jk})]^{1/2} \qquad [4.5]$$

与方程 4.1 到方程 4.4 一样，$B_j$ 代表第 $j$ 个职业类的虚拟变量的回归系数，$B_{jk}$ 表示第 $k$ 个自变量，在该例中，即为黑人与第 $j$ 个职业虚拟变量乘积项的回归系数。$t$ 检验可通过将表 4.4 中与黑人相关的系数的代入方程 4.5 计算得出。我们发现，所计算出的 $t$ 值没有一个是超过临界值±1.96 的，从而可得出结论，剔除了教育年限和工作任期的影响后，尽管职业定位对白人期望收入水平有显著的影响，却没有真正提高黑人的平均期望收入。[8]

至此，我们再回到图 3.1。我们将通过假设不同种族中的职业类别效应是不同的来开始本章的内容。就这张图，我们可以设想截距所体现的种族差异在各职业类别中是不同的。通过检验该假设，我们发现，结果确实如此。通过图 4.1与图 3.1 的对比，我们可以更明显地看出，不同职业类别的收入差异强度对黑人和白人确实不同。

与之前相比，表 4.5 列出了对图 3.1 和图 4.1 中 12 个组群的估计截距值。左边两列为通过模型 4 的估计值计算出来的截距(见图 3.1)，右边两列为通过模型 5 的估计值计算出来的截距(见图 4.1)。因为职业类别间种族的差异效应不同，所以通过模型 5 所得到的截距不仅修改了组群间的距离，还沿着 $Y$ 轴重新排列了组群。我们需要谨记在心的是，这些组群的期望收入均是建立在受访者的教育年限和工作任期均为 0 的情况下的。

从图 4.1 可以看出，白人和黑人在高级白领与技术工人之间的垂直距离都是一样的。对于这两个职业类别，代表技术工人的线比代表高级白领的线低2344美元。2344 美元为表 3.3 中列出的 $OCC_3$ 的回归系数，其估计了在年收入上，作为一个技术工人相对于高级白领所带来的影响的大小。然而，在图 4.1 中，高级白领与技术工人的比较描述了不同职业类别中的种族效应。对于白人，这段垂直距离为 2418 美元；对于黑人，这段距离仅为 92 美元。这些数据来自表 4.2 中模型 5 的回归结果，并可直接从表 4.4 中获得。

尽管图 3.1 与图 4.1 截距之间的距离不同，但是所有组群的教育年限和工作任期的偏斜率都是一样的。所有实线都是互相平行的，这表明图中所用的是

**表 4.5 种族和职业的截距差异**

| | 模型 4 | | 模型 5 | |
|---|---|---|---|---|
| | 白人 | 黑人 | 白人 | 黑人 |
| 高级白领 | 5761.1 | 4573.0 | 5794.8 | 2001.5 |
| 初级白领 | 3445.0 | 2256.9 | 3519.9 | 1227.8 |
| 技术工人 | 3417.4 | 2229.3 | 3376.4 | 1909.3 |
| 操作工人 | 2594.5 | 1406.4 | 2367.6 | 1559.1 |
| 服务业工作者 | 1842.6 | 654.5 | 1281.4 | 1016.1 |
| 劳工 | 2154.3 | 966.2 | 1592.0 | 1182.6 |

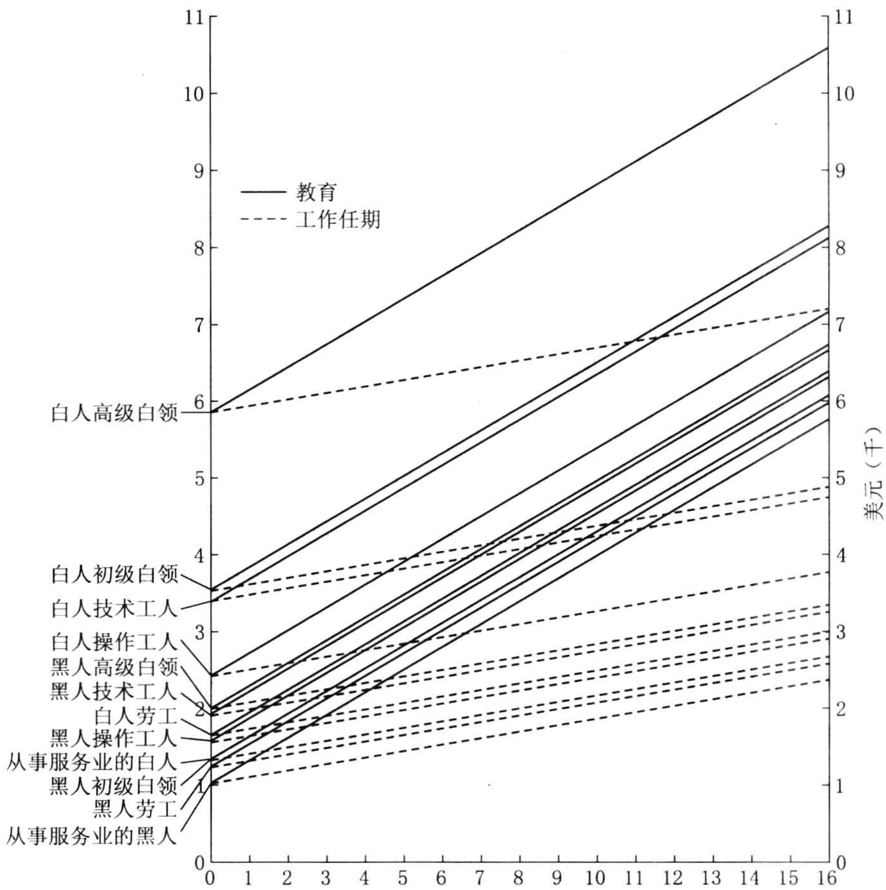



**图 4.1 模型 5 的回归结果**

教育的平均效应;所有虚线也是互相平行的,说明了工作任期也是由其平均效应表现的。本部分构建的最后一个模型对这点提出了质疑:对于所有组群,教育年限和工作任期的偏效应真的相同吗?我们可以假设教育和工作任期的种族效应,然后再检验教育年限和工作任期对黑人和白人是否都一样。尽管我们发现每增加一年教育年限或工作任期,所获得的收入增加值对黑人和白人是不同的,但是一些线仍然呈现平行状态。因为有6组线是与黑人相关的,其中一条线对应一个职业类别,还有6组线是与白人相关的,其中一条线同样对应一个职业类别。我们发现,教育对于种族的差异效应表明,与黑人工作者相关的实线是互相平行的,与白人工作者相关的实线也是互相平行的,但是这两个斜率可能不尽相同。[9]工作任期的也存在类似的情况。

为了检验假设 $\beta_{\text{EDUC(whites)}} = \beta_{\text{EDUC(blacks)}}$ 和 $\beta_{\text{TENURE(whites)}} = \beta_{\text{TENURE(blacks)}}$,我们构建了一个新模型来检验黑人和白人之间的差异效应。该模型基于模型5,只是在模型设定中又新加入了两个变量。由于要检验的是关系的变异性,因此我们仍会用到交互项。在该模型下,新增加的交互项分别为 EDUC 和 BLACK 的乘积及 TENURE 和 BLACK 的乘积,分别记做 BLEDUC 和 BLTEN。对于样本中的黑人,BLEDUC 的分布应该和 EDUC 的分布是一样的。然而,对于样本中的白人,其在 BLEDUC 中全被编码为 0;对于 BLTEN 与 TENURE,类似的情况仍然存在。该新模型可描述为:

$$\begin{aligned}\text{模型 6}: Y_i &= f(\text{种族},\text{职业},\text{教育},\text{工作任期}) \\ &= \beta_0 + \beta_1 \text{BLACK} + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 + \beta_4 \text{OCC}_4 + \beta_5 \text{OCC}_5 \\ &\quad + \beta_6 \text{OCC}_6 + \beta_7 \text{EDUC} + \beta_8 \text{TENURE} + \beta_9 \text{BLOCC}_2 \\ &\quad + \beta_{10} \text{BLOCC}_3 + \beta_{11} \text{BLOCC}_4 + \beta_{12} \text{BLOCC}_5 + \beta_{13} \text{BLOCC}_6 \\ &\quad + \beta_{14} \text{BLEDUC} + \beta_{15} \text{BLTENURE} + u_i\end{aligned}$$

模型回归结果在表 4.6 中列出。除了 EDUC 和 TENURE,其他与模型 5 中相同变量的回归系数的解释方法基本不变。有一点需明确的是,估计这些效应基于其不仅控制了其他自变量,还控制了种族在教育年限和工作任期上的差异效应。

**表 4.6 模型 6 的回归结果**

| | | 黑人的效应[a] |
|---|---|---|
| 常数 | 4962.5<br>(435.9) | |
| BLACK | −1667.3<br>(901.3) | |
| $OCC_2$ | −2155.4<br>(281.7) | −1068.8<br>(779.5) |
| $OCC_3$ | −2167.9<br>(242.5) | −718.4<br>(703.4) |
| $OCC_4$ | −3132.1<br>(268.8) | −1144.5<br>(672.7) |
| $OCC_5$ | −4281.2<br>(378.9) | −1605.1<br>(703.4) |
| $OCC_6$ | −3851.3<br>(411.0) | −1611.5<br>(703.7) |
| EDUC | 359.1<br>(29.4) | 186.3<br>(37.3) |
| TENURE | 80.3<br>(7.6) | 94.6<br>(13.1) |
| $BLOCC_2$ | 1086.5<br>(829.0) | |
| $BLOCC_3$ | 1449.5<br>(744.4) | |
| $BLOCC_4$ | 1987.6<br>(724.7) | |
| $BLOCC_5$ | 2676.1<br>(799.5) | |
| $BLOCC_6$ | 2239.7<br>(815.2) | |
| BLEDUC | −172.7<br>(47.5) | |
| BLTEN | 14.2<br>(15.2) | |
| $R^2$ | 0.32434 | |
| $F$ | 102.25 | |
| $R^2$ 的增量($R_3^2-R_1^2$) | 0.00296 | |
| 变化过的 $F$ | 7.01*** | |

注：括号里为回归系数的标准误。

a. 黑人的系数是由加和模型 6 回归系数得出的；其标准误通过 $[\text{var}(B_i)+\text{var}(B_j)+2\text{cov}(B_iB_j)]^{1/2}$ 计算得出。

*** 表示系数在 0.001 显著水平下显著。

尽管对于大多数变量，其系数估计和显著性检验都相同，但是相对于模型5，模型6有一些回归结果的变化需提及。尤其是虚拟变量黑人的系数减少了不止一半，而且其在0.05显著水平上也不显著了。另外，交互项$BLOCC_3$也明显变小了，而且在0.05水平下刚好显著。我们应该如何解释这些变化呢？

为了回答这个问题，我们必须看看新加入的这两个变量——教育和工作任期的交互项。BLTENURE的回归系数统计上并不显著，这表明在一个雇主的情况下，每增加一年的工作任期，其黑人或白人增加的收入是一样的。在这个发现下，我们回想从模型4到模型5，当第一次加入了乘积项后，从对$OCC_2$到$OCC_6$的系数的解释方法可以运用到该模型对TENURE的回归系数的解释中。在这里，TENURE的回归系数表明，为同一雇主工作每增加一年，白人工作者的收入每年会增加80.30美元。BLTENURE回归系数估计了黑人与白人的工作任期净差异为14.20美元。那么，我们就可以知道，黑人工作者为一个雇主工作每增加一年，其收入每年会增加94.50美元。然而很明显，BLTENURE的标准误非常小，这说明了该效应很弱。因此，我们可以得出如下结论：工作任期对黑人和白人的影响基本一样。

教育的情况就有所不同。EDUC的系数告诉我们，当控制了其他变量的影响之后，每增加一年的教育年限，白人的收入可增加359.10美元。BLEDUC回归系数表明，每增加一年的教育年限，黑人的收入会增加186.40美元（359.10美元－172.70美元）。BLEDUC的显著性检验表明，其在统计上是显著的。对于所有人群，若不考虑其他变量，那么，每增加一年的教育年限，黑人所增加的工资相对于白人会少一些。因为黑人和白人的教育净效应不同，所以模型5所用的黑人和白人的平均教育效应其实低估了每增加一年教育年限，白人的教育回报，同时高估了每增加一年教育年限，黑人的教育回报。

在描述有关种族—职业的交互项时，我们可以对这些影响效应问同样的问题。尤其当我们已经知道，工作任期对黑人的净效应与白人没有显著不同，但是教育的确不同，其对白人的净效应要大于黑人。我们不知道教育年限是否会显著影响黑人的期望收入水平。要回答这个问题，我们必须回到方程4.4，通过检验有关的回归系数的加和值来估计黑人的教育净效应。将相关系数代入方

程 4.4 后，我们发现：

$$t = 359.1 + (-172.7)/[(862.691) + (2253.569) + 2(-862.691)]^{1/2}$$
$$= 186.4/37.29 = 5.00$$

教育确实会影响白人和黑人的期望收入，然而，黑人的平均教育回报率比白人低。

对于之前提出的问题——如何考虑 BLACK 和 $BLOCC_3$的影响转移？现在我们可以提供这样一个解释：当如模型 5 一样，使白人和黑人在收入上获得相同的教育净效应后，黑人的回归系数告诉我们，在高级白领中，黑人有着非常显著的收入劣势；在控制了其他变量后，黑人高级白领平均比白人高级白领少挣4000美元。同样，对于技术工人与高级白领之间的期望收入净差异，白人(2418.40美元)也比黑人(92.20 美元)多很多。这些都可以从图 4.1 中看出。然而，一旦我们要估计模型 6 中种族对教育的差异效应时，就会发现在高级白领中，黑人和白人的期望收入净差异在双尾检验的 0.05 显著性水平下都不显著。尽管从模型 5 到模型 6，虚拟变量黑人的回归系数的标准差增加了，但是最重要的变化应该是系数本身的点估计。模型 6 中 BLACK 的回归系数比模型 5 中的一半还小，这个现象反映了黑人高级白领与白人高级白领的 $Y$ 截距的差异变小了。这点我们可以在图 4.2 中看出。但是，当我们沿着横坐标轴向高教育年限方向移动时，会发现黑人高级白领与白人高级白领间的距离变大，这说明白人高级白领相对于黑人高级白领的相对收入优势随着教育年限的提高而变大。从表 4.6 的回归系数我们可以算出，当教育年限为 0 时，白人高级白领与黑人高级白领的收入差异为 1667 美元；当教育年限变为 16 年时，白人高级白领的期望收入比黑人高级白领高 4430 美元。

当我们把注意力转移到 $BLOCC_3$时，会发现 $BLOCC_3$在模型 6 中的回归系数比模型 5 中的小，因此，我们不可以拒绝有关“当控制了模型中其他因素和教育的种族效应后，技术人员的净效应在黑人和白人之间相同”的零假设。但通过模型 5 和模型 6 的对比，有一点我们可以知道的是，白人高级白领与黑人高级白领之间的期望收入净差异部分可归因于获得比较高教育等级的黑人所挣得

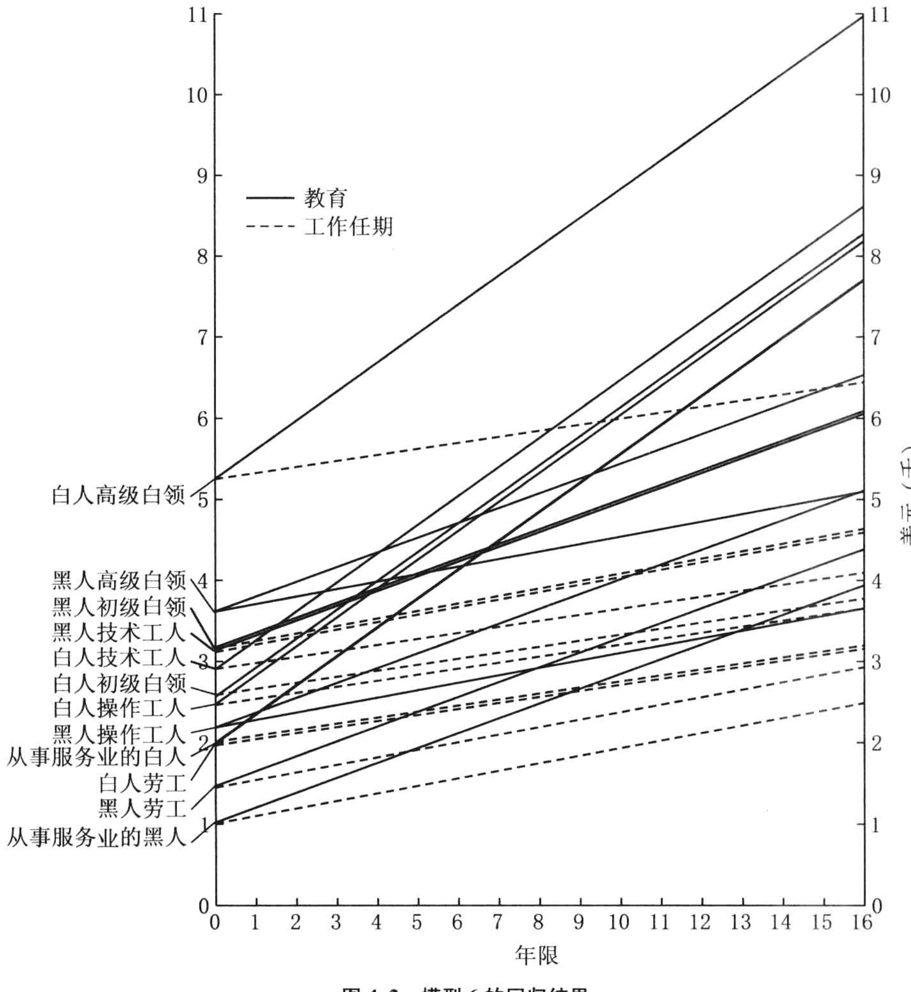



**图 4.2　模型 6 的回归结果**

的收入，总是不如同等条件下的白人多。如果额外的教育年限带来的收入差异被允许，比如，一旦我们可以承认额外的教育年限所引起的收入增加在白人中比黑人多，我们就可以解释为什么黑人高级白领获得的期望收入比白人高级白领少，因为他们额外教育的回报率相对白人而言很低。我们不想把最初的观测结果——黑人高级白领相比白人高级白领，有收入劣势——视为无效，相反，根据黑人的收入劣势，模型 6 的结果提出了一个可能的解释。另外，我们知道，在

技术工人中,那些最初看上去由种族差异导致的收入差异效应,其实部分也归因于在该职业类别中,黑人相对于白人所累积的教育回报差异。

## 对各组群分别进行回归

我们从研究年纪较大的黑人工作者与白人工作者的期望收入差异开始,现在,需要通过控制职业、教育年限、工作任期来研究,进而一步步扩展模型设定,把有关白人与黑人中的额外变量的差异效应检验也包括在内。最后一步可能会引起读者的疑问,在允许自变量在每个组群中有所不同的情况下,为什么我们会用模型 6 对整个样本进行估计,而不是对每个组群分别进行回归估计呢?为什么不把样本分成黑人和白人两组,用期望收入对职业虚拟变量、教育年限和工作任期对每个组群分别进行回归?事实上,若是检验假设和标准 OLS 假设都可以通过恰当的统计过程得到满足,那么,这些方法都是等价的。

为了构建一个含有交互项的全样本模型,我们要注意以下 6 点:

第一,在制定的乘积项或交互项缺失时,自变量的系数告诉我们的是一个"平均效应",而当其他自变量也包含在规范里时,则为"平均偏效应"。

第二,当把乘积项拓展到规范中时,我们可以通过比较两个模型的 $R^2$ 值来确定是否要用各组群的平均效应来提高模型拟合度。如果 $R^2$ 的增量是由于加入了乘积项后而变得足够大,我们就可以拒绝零假设(各组群的效应是相同的)。

第三,当我们用模型 6 对全样本进行估计时,从 $OCC_2$ 到 $OCC_6$ 的回归系数的 $t$ 检验可以测量白人工作者的职业净效应,而 BLACK 的回归系数的 $t$ 检验则可以解释非洲裔美国人在高级白领中的期望收入净效应。

第四,为了检验一个自变量的效应是否对黑人工作者显著,我们必须构建两回归系数的加和的 $t$ 检验。

第五,为了检验两个自变量的效应是不是显著不同,例如,检验操作工人是不是和服务业工作者不同,我们必须构建两回归系数差异的 $t$ 检验。

第六,要对乘积项进行 $t$ 检验,我们需要确定解释变量的效应是否因种族差

异而不同。

通过对每个组群分别进行回归，例如，先对黑人进行回归，再对白人进行回归，我们就可以自动估计出不同组群的效应。换句话说，对每个组群，如果要知道一个指定自变量是否有显著效应，就必须考虑上述第四点是否必要。然而，如果估计不同组群的回归的目的是评定组群差异效应的显著程度，那么，该差异效应检验就非常必要了。

当缺乏详尽的检验时，研究者可能会陷入两难。想象一下，当要分析政治激进主义是年龄的函数时，我们期望得到的是年龄和政治激进主义的关系随教育程度的不同而不同，尤其是当我们假设年龄和政治激进主义的关系是否与大学毕业有关时。从大学毕业组与非大学毕业组分别随机选取 500 个样本，然后分别对其进行回归，我们可能会发现一个明显的年龄期望效应“差异”。例如，假设年龄对大学毕业生的影响为－0.16，而对非大学毕业生的影响为－0.32。那么，因为大学毕业生的回归系数是非大学毕业生的一半，我们就可以说大学毕业生的影响比那些没有接受大学教育的人小吗？大多数读者在此可能会意识到这么说有风险。如果该研究的目的是从样本推及整体，那么同时考虑点估计和误差是非常必要的。当要评估回归模型中自变量影响的重要性时，就不是简单的系数强度的问题了，而是系数强度相对于标准差的问题。当要描述有关相对影响的强度，即这些影响是不是相等时，也需要考虑该问题。估计回归系数的差异强度必须相对于差异的标准差。在这种情况下，由于信息非常有限，我们很难判断在受教育程度不同的组群里，年龄对政治激进主义的相对影响。

**表 4.7　对组群分别回归的结果**

| | 黑　人 | 白　人 |
|---|---|---|
| 常数 | 3295.2<br>(416.0) | 4962.5<br>(494.9) |
| $OCC_2$ | －1068.8<br>(411.2) | －2155.4<br>(319.8) |
| $OCC_3$ | －718.4<br>(371.1) | －2167.9<br>(275.3) |

续表

| | 黑　人 | 白　人 |
|---|---|---|
| $OCC_4$ | −1144.5<br>(354.9) | −3132.1<br>(305.2) |
| $OCC_5$ | −1605.1<br>(371.2) | −4281.2<br>(430.2) |
| $OCC_6$ | −1611.5<br>(371.2) | −3851.3<br>(466.6) |
| EDUC | 186.3<br>(19.7) | 359.1<br>(33.3) |
| TENURE | 94.5<br>(6.9) | 80.3<br>(8.6) |
| $R^2$ | 0.31887 | 0.24450 |
| $F$ | 61.06 | 105.50 |
| RSS/$(n-k-1)$ | 4046652.4 | 18754709.2 |
| $N$ | 921 | 2290 |

注:括号里为回归系数的标准误。

现在我们假设,除了回归系数估计,我们已经知道了这些估计的标准差:回归系数−0.16 的标准差为 0.11,系数 0.32 的标准差为 0.14。那么我们会发现,大学毕业生组的影响在常用的显著性水平下并不显著。[10]此时,是不是就可以下结论说,年龄的效应在非大学毕业生组更大呢?答案当然是否定的。尽管我们已经证明在非大学毕业生组中,年龄是估计政治激进程度的一个重要预测因子,但是我们并没有对该差异效应进行进一步的推理。如果研究问题是:两种影响是不是相等?或者说 $H_0:\beta_{CG}-\beta_{NCG}=0$ 是否成立(其中,$\beta_{CG}$是年龄在大学毕业生组的影响,$\beta_{NCG}$为年龄在非大学毕业生组的影响)?那么此时,就不再是 $\beta_{CG}$或 $\beta_{NCG}$本身相对于我们的零假设是不是显著,比较合适的估计应该是用 $\beta_{CG}-\beta_{NCG}$。该统计强度必须是相对于 $\beta_{CG}-\beta_{NCG}$ 的标准差来估计的。当回归系数都来自同一个方程时,我们可以用方程 4.5 来进行检验。但是,如果系数是从不同的回归中估计出来的,该检验的定义就会有些不同。[11]

要说清楚这些问题,我们必须先回到收入数据。然后对白人工作者和黑人工作者分别进行回归,其回归结果可在表 4.7 中看到。由于回归系数是在不同

的样本中估计出来的，那么，所有 $B$（如 $B_1$、$B_2$、$B_3$ 等）差异影响估计必然是不相关的。那就是说，协方差的估计为 0。因此，方程中的标准差变为根号下方差的和。然而，这个不同组群的方差系数部分是基于整体方差的组群估计。总体方差因为总体方差的每两个估计只是建立在“部分”样本上的，所以其用的是残差平方和的不同部分。这就表明了通过合并两个组群的信息计算总体方差的合并估计值的必要性（Kmenta，1986）。另外，由于组群的大小可能不同（例如，白人的数量可能比黑人多两倍以上），这时，该合并估计值必须通过合适的自由度来对每个组群估计加权（Long & Miethe，1988）。

假设各组群的方差是相等的（也就是我们所说的方差的同质性），那么，计算总体方差的合并估计值的方程为：

$$s_{\text{pooled}}^2 = \frac{(n_1 - k_1 - 1)s_1^2 + (n_2 - k_2 - 1)s_2^2}{N - (k_1 + k_2 + 2)} \quad [4.6]$$

其中，$n_1$ 和 $n_2$ 是组群中的事件数，$N = n_1 + n_2$；$k_1$ 和 $k_2$ 为每个组群中包含的自变量的数目；$s_1^2$ 和 $s_2^2$ 是组群各自回归出的平均残差平方和。[12] 用来检验组群回归系数差异的 $t$ 检验公式为：

$$t = \frac{B_1 - B_2}{s_{\text{pooled}}\left(\frac{s_{B_1}^2}{s_1^2} + \frac{s_{B_2}^2}{s_2^2}\right)^{1/2}} \quad [4.7]$$

其中，$s_{B_1}^2$ 和 $s_{B_2}^2$ 分别为 $B_1$ 和 $B_2$ 的方差，$s_1^2$ 和 $s_2^2$ 与上式相同。通过执行上述 $t$ 检验，我们即可重现在全样本回归模型中的交互项的 $t$ 检验。例如，将表 4.7 中的回归结果代入方程 4.6 及方程 4.7，从而检验对黑人和白人教育的影响是不是都一样。那么，为了计算总体方差的合并估计，我们有：

$$s_{\text{pooled}}^2 = \frac{(921 - 8)(4046652) + (2290 - 8)(18754709)}{3211 - 16}$$

$$= 14551750$$

将其带入 $t$ 检验方程后，我们发现：

$$t = \frac{186.3 - 359.1}{3814.7\left(\frac{386.8}{4046652} + \frac{1108.9}{18754709}\right)^{1/2}} = \frac{-172.7}{47.5} = -3.6$$

通过 $t$ 统计量分子、分母的比较，连同表 4.6 里所有回归系数和 BLEDUC 的标准差，我们可以对这两个过程的等价性看得更加清楚。然而，在证明该等价性时，我们仅简单地把相关系数带入方程 4.6 和方程 4.7，而没有检查是否满足了隐藏在 $t$ 检验背后的假设。事实上，直到现在，所有的讨论都忽略了是否符合 OLS 假设的问题。在构建一个含有二进制编码的虚拟变量回归系数的解释过程中，我们一直坚持 OLS 假设是没有争议的，不管事实如何。现在是时候纠正这种错误了。确实，由于当前我们把重心放在推理检验上，这就需要我们更仔细地检验这些假设，特别是在已经有了先例后，在对有关从全样本或者组群回归中出现的差异效应下任何结论前，我们必须检验方差的同质性（或称“同方差性”）。

但是，尽管是否符合 OLS 假设非常重要，但从表 4.7 和表 4.6 的比较结果来看，这样做还是会有问题。表 4.6 中列出的黑人职业净效应的显著性检验是基于模型 6 的全样本估计，当其分别对黑人与白人回归后，无法得到重现。虽然对回归系数的估计没有分别，但是标准差却明显不同。当对各组群分别进行回归估计时，相比表 4.6 中的结果，黑人的标准差变小了，而白人的标准差变大了。那么，我们如何解释这种不一致性呢？

## 处理异方差性

我们已经根据隐含的假设（收入水平和收入结构的决定因素与种族有关并随种族不同而变化）发展了模型。换句话说，我们开始只是识别黑人和白人的收入水平总差异，之后发展到可以解决解释变量差异效应的问题，比如职业虚拟变量和与教育年限、工作任期有关的交互项等。然而，这些检验都存在潜在问题，因为其假设是基于两组的同方差性。

尽管从模型 6 中用全样本数据获得的回归系数估计和分别回归组群数据得

出的估计一样，但是由全样本回归得来的总体方差 RSS/($N-k-1$) 只有在同方差性的假设被满足后，才基本上与分别回归组群得来的方差合并估计值相等。由组群回归带来的平均残差平方和对于黑人是 4046652，而对白人为 18754709。模型 6 的全样本估计用了所有的观测值，相对于组群回归估计每个组群只估计了其中一部分的观测，在该例中，只有 921 个黑人或者只有 2290 个白人，全样本估计是建立在最多的信息上的。但是，只要我们假设 $u_i$ 对所有的 $X_{ki}$ 都是个常数，那么从全样本数据得来的估计即对总体方差的无偏估计。但是对于本例，"$u_i$的方差不会因组的不同而不同"这一假设有效吗？要回答这个问题，必须更仔细地研究我们的回归结果。

同方差性的假设指出，给定特定值的自变量干扰项的条件方差为一个常数，即 $\sigma^2$。对于比较不同组群的研究问题，检验其异方差性是很有必要的。许多异方差检验都要求研究者检验 OLS 估计的残差平方，即 $e_i^2$。基本的统计教材也会为该问题及其检验过程提供比较概述性的讨论(Gujarati，1988；Johnston，1984)。根据我们的侧重点，我们将会把注意力主要集中到两个检验和一个讨论上。

在此例中，我们假设方差的大小是种族的函数，即方差在不同的种族中是不同的。一个直接的检验是可以比较从黑人和白人的分别回归中得到的平均残差平方和。如果同方差性存在，那么两组群在估计回归平面的误差方差上也应该相同。因此，我们可以提出零假设，即这些方差都是相等的，其又可表达为，$H_0:\sigma_1^2=\sigma_2^2$，同时可用平均残差和作为这些参数的估计。如果我们构建了一个有关方差的比率，那么当比率为 1 时，方差相等；当比率偏离 1 时，等方差的假设就站不住脚了。如果假设 $u_i$ 服从正态分布且同方差性的假设成立，那么该比率遵循 $F$ 分布。因此，我们可以构造如下的检验：

$$F_{n_1-k_1-1,\ n_2-k_2-1}=\frac{\text{RSS}_1/n_1-k_1-1}{\text{RSS}_2/n_2-k_2-1} \qquad [4.8]$$

其中，分子由较大方差的组群回归得出，而分母由较小方差的组群回归得出。在该情形下，$k_1$ 和 $k_2$(回归中所包含的自变量个数)是相等的，因其模型设定相同。将表 4.7 的结果代入，我们有：

$$F_{9\ 132\ 282} = 18754709.2/4046652.4 = 4.63$$

对于该样本量大小的组群，$F$ 值在 0.001 显著水平下为 4.63，这表明我们要拒绝同方差的零假设，而倾向于异方差性。对于模型 6，其回归系数的估计仍然是无偏的。然而，在异方差下，$t$ 检验是不准确的。而且，在组群间的异方差条件下阐述解释变量的组群差异效应非常复杂，因为影响差异的回归检验不清楚。我们并不知道导致该检验结果的是组群间的差异效应，还是组群间的方差差异。

当方差不等时，检验正态分布的均值的等价性的问题即著名的Behrens-Fisher 问题(Amemiya, 1986:36)。针对该问题，已经有了很多解决方法。这些方法主要通过越过不等方差问题或重新计算检验统计量的分布来调整偏差，其大多或依赖某种数据转换，或重新权重。在本例中，异方差性的问题似乎比较容易解决。

最初，我们会在传统回归模型设定的基础上着手，这样，解释变量的备择组合就有被假定的可能性，但是我们不会质疑用来检验这些关系的模型设定。虽然已经用 INCOME(以工资和奖金挣来的美元数衡量)作为因变量，但是实际所用的模型设定可能并不是最佳选择。例如，如果工资呈对数的正态分布(该断言已经为大量的经济学文献所支持)，那么异方差性可能就会在模型设定的误差中产生。

## 解释半对数方程的虚拟变量

我们通常使用自变量和因变量原来的度量来解释回归方程。根据此惯例，我们保留了读者已非常熟悉的回归系数解释方法，即期望值 $Y$ 随 $X$ 的单位变化而变化。有时，函数规范会要求自变量或因变量，或者两者一起变换。研究文献中最常见的变换就是对数变换。当变量呈高度偏斜分布时，这种变换就尤其有用。[13]尽管对数变换通常是处理异方差性的补救方法(Gujarati, 1988; Maddala, 1992)，但在本例中，将其转变成收入的自然对数是为了澄清一个对收入分布

的理解问题，即 5000 美元与 10000 美元的收入差异和 50000 美元与 55000 美元的收入差异的意义是不同的。而在初始的收入分布度量中，不管该增量在分布中是如何计算的，5000 美元的差异就是 5000 美元的差异，它始终只是一个增量而已。相反，如果把增量 5000 美元视为比例项，我们就可以看出，在初始5000美元的基础上再增加 5000 美元，相当于增加了 100%，而在初始 50000 美元的基础上增加 5000 美元，相当于只增加了 10%。那么，如果要说出与 50000 美元成比例的同等效力，则应该是在其基础上增加到100000美元，或者说增加 50000 美元，这样才增加了 100%。通过收入的对数变化，我们可以用比例项来表示自变量和因变量之间的关系。在一个半对数模型中，只能通过将因变量或者自变量转换为对数形式来实现；在本例中，我们所用的是因变量收入的对数变换模式，而保持自变量的度量标准不变。

我们可以将 $Y$ 的半对数模型定义如下：

$$
\begin{aligned}
\text{模型 7: } \ln(Y_i) &= f(\text{种族，职业，教育，工作任期}) \\
&= \beta_0 + \beta_1 \mathrm{BLACK} + \beta_2 \mathrm{OCC}_2 + \beta_3 \mathrm{OCC}_3 + \beta_4 \mathrm{OCC}_4 \\
&\quad + \beta_5 \mathrm{OCC}_5 + \beta_6 \mathrm{OCC}_6 + \beta_7 \mathrm{EDUC} + \beta_8 \mathrm{TENURE} \\
&\quad + \beta_9 \mathrm{BLOCC}_2 + \beta_{10} \mathrm{BLOCC}_3 + \beta_{11} \mathrm{BLOCC}_4 \\
&\quad + \beta_{12} \mathrm{BLOCC}_5 + \beta_{13} \mathrm{BLOCC}_6 + \beta_{14} \mathrm{BLEDUC} \\
&\quad + \beta_{15} \mathrm{BLTENURE} + u_i
\end{aligned}
$$

当 $X_{ki}$ 为一个连续变量时，我们将其回归系数解释为在给定的 $X$ 的绝对变化下，$Y$ 的相对变化。例如，TENURE 变化一年，$Y$ 的成比例变化是多少。如果我们用 100 乘以 $\beta_8$，那么，我们得到的是 $X$ 一个单位的绝对变化所带来的收入的百分比变化。例如，如果 $B_8$ 为 0.014，我们可以说每增加一年的工作任期，所预测的 $Y$ 的变化为 1.4%。尽管该解释在自变量为连续度量时是有效的，但是 Halvorsen 和 Palmquist(1980)告诉我们，对于虚拟变量，其回归系数不可以这样解释。

假设 $X_{ki}$ 是一个虚拟变量，由于虚拟变量用的是离散编码，如 0 和 1，那么就不可以通过回归系数来定义斜率。因此，对于虚拟变量，我们不能将其回归系数作为因变量的衍生物。另外，虚拟变量的回归系数通过因变量的单位变化捕

捉的是指定组群与参照组之间的均值差异。当 $Y^* = \ln Y$ 时，半对数模型设定中的虚拟变量的回归系数表达了 $\ln Y$ 变化的单位数。就如 Halvorsen 和 Palmquist(1980)所说的，半对数回归中的虚拟变量系数等于：

$$B_1 = \ln \frac{1 + \hat{Y}_1 - \hat{Y}_{\text{ref}}}{\hat{Y}_{\text{ref}}} \qquad [4.9]$$

其中，$\hat{Y}_1$ 为被编码成 1 的组群的 $\hat{Y}$ 期望值，$\hat{Y}_{\text{ref}}$ 为参照组的 $Y$ 的期望值。为了用 $Y$ 的原始度量标准而不是相对于其对数变换模式的分布找到虚拟变量对 $Y$ 的百分比效应，我们就需要用对数的反函数，即指数形式或者反对数形式。此时，编码为 1 的组群(而不是参照组)的百分比差异可表示为：

$$100[\exp(B_1) - 1] \qquad [4.10]$$

因此，若虚拟变量的系数，如黑人的回归系数 $B_1$ 为－0.632，那么我们会发现，用 1 减去以 e 为底、－0.632 为指数的值为－0.468。该指定组的期望值 $Y$ 在本例中为 0.468(46.8%)，比参照组白人的期望值低。

表 4.8 的左边列出了模型 7 的全样本估计结果；右边是对黑人和白人分别进行回归的结果。

我们第一个任务是观察现在是否满足同方差的假设。将相关系数代入方程 4.8 后，我们有：

$$F_{9\ 132\ 282} = 0.27774/0.21983 = 1.263$$

根据所得的 $F$ 值，我们无法拒绝同方差的零假设，因此，现在可以着手检验黑人和白人之间的差异效应了。

在总结重要的结论之前，我们先来比较一下表 4.8 里两个回归模式的结果。首先，我们可以发现，表的两部分的系数估计是相同的。其次，如果再仔细观察会发现，不同回归模式所得到的标准差也非常接近。第三，通过这两个模式的回归检验，我们可以得到相同的结论：对于白人，在控制教育和工作任期的情况下，高级白领有着非常显著的收入优势，同时，教育年限和工作任期对期望收入有正的净效应；对于高级白领，黑人没有白人那样的收入优势，事实上，当控制了模型中的其他因素时，高级白领的期望收入比劳工多很多，但是其他的职业

**表 4.8 所有解释变量和交互项对 LN(INCOME)的效应**

| | 全样本回归 | | 组群样本回归 | |
|---|---|---|---|---|
| | 模 型 | 黑 人[a] | 白 人 | 黑 人 |
| 常数 | 8.353<br>(0.056) | | 8.353<br>(0.054) | 7.720<br>(0.109) |
| BLACK | −0.632<br>(0.115) | | | |
| $OCC_2$ | −0.244<br>(0.036) | −0.029<br>(0.099) | −0.244<br>(0.035) | −0.029<br>(0.108) |
| $OCC_3$ | −0.174<br>(0.031) | 0.056<br>(0.090) | −0.174<br>(0.030) | 0.056<br>(0.097) |
| $OCC_4$ | −0.328<br>(0.035) | −0.056<br>(0.086) | −0.328<br>(0.033) | −0.056<br>(0.093) |
| $OCC_5$ | −0.585<br>(0.049) | −0.105<br>(0.090) | −0.585<br>(0.047) | −0.105<br>(0.097) |
| $OCC_6$ | −0.510<br>(0.053) | −0.209<br>(0.090) | −0.510<br>(0.051) | −0.209<br>(0.097) |
| EDUC | 0.049<br>(0.004) | 0.043<br>(0.005) | 0.049<br>(0.004) | 0.043<br>(0.005) |
| TENURE | 0.014<br>(0.001) | 0.027<br>(0.002) | 0.014<br>(0.001) | 0.027<br>(0.002) |
| $BLOCC_2$ | 0.215<br>(0.106) | | | |
| $BLOCC_3$ | 0.230<br>(0.095) | | | |
| $BLOCC_4$ | 0.272<br>(0.093) | | | |
| $BLOCC_5$ | 0.480<br>(0.102) | | | |
| $BLOCC_6$ | 0.301<br>(0.104) | | | |
| BLEDUC | −0.006<br>(0.006) | | | |
| BLTEN | 0.013<br>(0.002) | | | |
| RSS的平均值 | 0.23654 | | 0.21983 | 0.27774 |
| $R^2$ | 0.42489 | | 0.32987 | 0.30082 |

注:括号里为回归系数的标准误。

a. 黑人的系数是由加和模型 6 的回归系数得出的;其标准误通过 $[\mathrm{var}(B_i)+\mathrm{var}(B_j)+2\mathrm{cov}(B_iB_j)]^{1/2}$ 计算得出。

差异却没有那么显著。教育年限所带来的收入净效应对黑人和白人来说没什么差别,然而,工作任期所带来的效应对黑人比对白人大。此外,种族和职业的交互项告诉我们,职业的净效应因种族的不同而不同,职业差异导致的期望收入差异在黑人中表现甚小。

我们可以通过描述黑人在各职业中的净效应来总结这些结果:黑人高级白领相对于白人高级白领有明显收入劣势,这种劣势基本上在除了服务业工作者以外的所有职业类别中存在。在服务业工作者中,控制了教育和工作任期后,黑人和白人之间的期望收入差异并不显著。[14]

## 检验两组以上的异方差性

我们可以找到许多关于检验异方差性的方法及文献。例如,Goldfeld-Quandt 检验,其在观测量不是很大之时是一个比较合适而且常用的方法。然而,此方法需要把观测分成两组(Goldfeld & Quandt, 1972; Gujarati, 1988)。另一个常被引用的检验过程是 Glejser(1969)提出的,他认为要检验一系列回归,就要证明$|e_i|$是模型中每个自变量的函数(其局限性可参见 Gujarati, 1988; Maddala, 1992)。在该例中,我们先用模型 6 的回归结果来计算 $e_i$,然后用$|e_i|$对 BLACK 进行回归。此方程的 $F$ 检验是为了说明是否要拒绝误差项的同方差性。从两组扩展到 $j$ 组要求$|e_i|$对 $j-1$ 个虚拟变量进行回归,此时,$F$ 检验可决定是否拒绝每组方差都是相同的假设。

异方差性的检验常常在判断已经被估计了的回归模型时被使用。当遇到这个问题时,研究者对决定是否在开始进行回归前就检验方差同质性比较感兴趣。最近常用的一个同方差性检验的比较分析告诉我们,这些检验的检定力和稳健性的方差非常大(Conover, Johnson & Johnson, 1981)。这些检验的一个共同局限就是它们对非正态分布的结果非常敏感。Levene(1960)提出了一个在比较分析中运用的较好检验,该检验的结构其实与 Glejser 的异方差性检验非常类似。Levene 认为,用单因素方差分析分析绝对偏差值时,如果用偏差的中位数代替偏差均值,其稳健性检验会明显提高。为了进行该检验,研究者必

须在一开始就计算好$|Y_{ij}-\widetilde{Y}_j|$，其中，$\widetilde{Y}_j$ 表示第 $j$ 组中位数。因为单因素方差分析等价于用一个虚拟变量和其他 $j-1$ 个虚拟变量进行回归，所以我们就可以估计回归方程：

$$|Y_{ij}-\widetilde{Y}_j|=B_0+B_1D_1+\cdots+B_{j-1}D_{j-1}+e_i$$

该方程的 $F$ 检验决定了同方差性零假设是否应该被拒绝。

## 用非独立检验进行多重比较的方法

通过讨论有关估计组群差异的方法，我们找到了另一个值得讨论的话题——多重比较的问题。该问题是同时统计学推论中的一个论点（Miller，1966），源于对单系列的估计进行多重组对比。我们要进行比较的数目越大，至少有一个比较是显著的可能性也越大。现在对多重比较进行显著性检验的合适方法仍然是一个争论的焦点。这里，我会简单介绍两种方法——Bonferroni检验和 Fisher 的 $t$ 保护方法（Darlington，1990：249—275）。

在本例中，我们检验了不同职业类别的期望收入差异。把高级白领作为参照组，表明我们会直接比较 5 对职业类别，即其他 5 类职业相对于参照组的差异。然而，我们也引入了 $t$ 检验，它可以检验回归系数之间的差异。我们知道，所有可能的对比数目共有 $j(j-1)$ 个，其中 $j$ 表示类别的数目。然而，若比较的先后顺序无关紧要，可能的对比数目会减少一半。由此，对于 6 个职业组，我们可以生成 15 个可能的成对比较。用这些方法来阐述该问题，需要我们明确这些可能的比较是否相互独立。对于该例，高级白领与初级白领的比较和高级白领与技术工人的比较不是相互独立的，因为对收入较高的高级白领的概率选择可能会影响这两对比较结果，所以高级白领与初级白领、技术工人与操作工人的对比才是独立的比较，或者说是正交的比较。然而，对这些对比的显著性检验不是完全独立的。因为在计算这两对比较的标准差时，我们用的都是总体方差的估计值，即 RSS/$(N-k-1)$，所以该估计的随机波动会影响两个检验的 $t$ 值。

Ryan(1960)发展了对 Bonferroni 不等式和独立检验的应用,他证明了Boferroni不等式为大多数修正过的独立检验的显著性水平提供了略为保守的估计。该不等式的目的是为了根据多重比较的实际情况,提供一个修正后的显著性水平。Bonferroni 检验类似于修正显著性水平(CSL),通过把检测得出的所有结果的最显著 $t$ 值所对应的概率 $p$ 与被检验的结果个数(即 Bonferroni 校正因子或者 BCF)相乘,从而可以计算出一个修正显著性水平。因此,若 15 个结果中最显著的 $t$ 值对应的 $p$ 值为 0.003,那么对该估计修正显著性水平即 $CSL = 15(0.003) = 0.045$。其中,15 就代表 BCF。如果要用第二小的 $p$ 值来估计系数的显著性,我们就应该将该 $p$ 值乘以 14。该方法被称为"压条法",这个过程一直到找到第一个非显著结果才会停止。

Dunn(1961)告诉我们,对于双尾检验,表达式$[1-(1-CSL)^{BCF}]$可为我们提供一个修正显著性水平的上限。然而,Bonferroni 检验有时太过保守,Dunn 检验有时又显得太过自由。如果想更好地理解 Bonferroni 不等式的用法,我们必须考虑两个极端情况(Darlington, 1990)。第一个情况为两个检验在$-1$上相关。例如,如果我们对 $B_k$ 进行双尾检验,则要检验两个零假设:$\beta_k \geqslant 0$ 和 $\beta_k \leqslant 0$。这两个检验在$-1$上相关,因为要拒绝第一个零假设,则必须排除第二个零假设。用 Bonferroni 不等式,我们可以说,至少一个零假设在0.025显著性水平被拒绝的可能性不会比 0.05 大,即 $2 \times (0.025)$。在该情形下,$p$ 值和修正显著性水平 0.05(即一个双尾检验零假设 $\beta_k = 0$)的显著性水平一样大。因此,在该情况下,Boferroni 不等式并不保守。第二个情况即两个检验在$+1$上相关。假设我们有一个包含 $j$ 个类别的变量,我们每次都用第一个类别去和其他任何一个类别比较,这样就可以产生 $j-1$ 个比较。该情况在我们有一个 $j$ 类的定性变量并且在回归模型里含有根据该变量产生的 $j-1$ 个虚拟变量时便会发生。如果所有除去参照组的类别都有无限的样本量和同样的均值,那么所有比较的 $t$ 值也会一样,此时,拒绝一个事件的零假设就意味着其他的零假设也要被拒绝。在这种情况下,修正显著性水平就等于我们观测到的 $p$ 值。因此,如果我们要用 Bonferroni 不定式,我们就要高估修正显著性水平。因此,Bonferroni 公式的不准确性与各检验之间的相关性有关。各检验间的正

相关越大，误差就越大。

相比之下，Fisher 的方法就更开放一些。运用该方法时，研究者进行 $F$ 检验来检验各类别之间没有不同的零假设是否要被拒绝。如果 $F$ 检验结果显著，那么研究者就可以进行任意类别的比较，因为包含在这些对比中的 $t$ 检验已被从 $F$ 检验得出的显著性结果证实了。在讨论各种模型的回归结果中，第一步总是检验当加入一系列虚拟变量来代表一个定性特征（如职业的一系列虚拟变量）或者一系列交互项（如种族和职业的乘积项）后，所出现的 $R^2$ 的增量的统计显著性。在通过 $F$ 检验建立统计显著性后，我们就可以用 Fisher 方法来进行多重比较了。

# 第 5 章 | 可替代虚拟变量编码方案

迄今我们所接触到的虚拟变量大都采用二进制编码并指定单一参照组，其实还有很多其他的编码方案也是可行的。比如，效果编码和对比编码这两种替代性的方案。这两种方法要求我们用 $j-1$ 个虚拟变量来表示具有 $j$ 个类别的名义变量。

## 效果编码虚拟变量

如第 2 章提到的，有些研究者倾向于选择一个中间类别作为参照组，而不是直接用一组按序数分布的极端类别，这样的选择可以解释为，通过建立组群比较来模拟指定类别和所有样本的"平均值"的差异。如果有人想对组群和样本平均进行比较，效果编码的解释结构将比二进制编码更方便。

**表 5.1 虚拟变量的效果编码和对比编码**

| 职业类别 | 效果编码 | | | | |
|---|---|---|---|---|---|
| | $E_1$ | $E_2$ | $E_3$ | $E_4$ | $E_5$ |
| 高级白领 | −1 | −1 | −1 | −1 | −1 |
| 初级白领 | 1 | 0 | 0 | 0 | 0 |
| 技术工人 | 0 | 1 | 0 | 0 | 0 |
| 操作工人 | 0 | 0 | 1 | 0 | 0 |
| 服务业工作者 | 0 | 0 | 0 | 1 | 0 |
| 劳工 | 0 | 0 | 0 | 0 | 1 |
| 高级白领 | 0.5 | 1 | 0 | 0 | 0 |
| 初级白领 | 0.5 | −1 | 0 | 0 | 0 |

续表

| 对比编码 | 效果编码 | | | | |
|---|---|---|---|---|---|
| | $C_5$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
| 技术工人 | −0.25 | 0 | 0.5 | 1 | 0 |
| 操作工人 | −0.25 | 0 | 0.5 | −1 | 0 |
| 服务业工作者 | −0.25 | 0 | −0.5 | 0 | 1 |
| 劳工 | −0.25 | 0 | −0.5 | 0 | −1 |

为了方便用二进制编码和其他编码方案进行比较，我们将继续将收入作为因变量，将种族和职业作为名义自变量。表 5.1 举例说明了我们如何分别利用效果编码和二进制编码虚拟变量来捕捉种族和职业类别的信息。表格的上半部分描述了通过效果编码产生的 5 个虚拟变量，其保留了高级白领作为参照组。然而，我们知道在二进制虚拟变量编码中，参照组通常被编码为 0，而对于通过效果编码产生虚拟变量的参照组，习惯编码为 −1。每一个虚拟变量所捕捉的职业类别对比即参照组和编码为 1 的组。在本例中，$E_1$ 为高级白领和初级白领的对比；$E_2$ 为高级白领和技术工人的对比；$E_3$ 为高级白领和操作工人的对比等等。如果需要比较的组群样本大小一样，编码为 0 的组就不会影响比较结果。但是，如果组群的样本大小不一致（这是常见的例子），编码为 0 的组群所带来的影响便会出现，尽管其很小。实际上，编码为 0 的组群的影响随着所有编码为 0 的观测值的偏离样本均值而增大(Cohen & Cohen, 1983)。

表 5.2 记录了效果编码虚拟变量和收入的零阶相关性、均值及标准偏差。除了职业类别的虚拟变量，还有一个种族虚拟变量 ERACE，其为白人时编码为 1，为黑人时编码为 0。二进制编码虚拟变量均值等同于指定组群中的事件比例，而效果编码虚拟变量的均值则指出了参照组（编码为 −1）和其他编码为 1 的组群的大小差异。实际上，均值即 $(n_j - n_{\mathrm{ref}})/N$。例如，$E_1$ 将 644 个高级白领编码为 −1，337 个初级白领编码为 1，其他为 0。那么，$E_1$ 的均值为 $(337 - 644)/3211 = -0.096$。负号表示参照组比编码为 1 的组群具有更多的观测值；强度表示这个差异的大小相对于总体样本的大小。这组均值指出，高级白领的观测值数量超过了初级白领、服务业工作者和劳工，但少于技术工人和操作工

表 5.2 虚拟变量的均值、标准差和相关性

| | 效果编码的虚拟变量 | | | | | |
|---|---|---|---|---|---|---|
| | $E_1$ | $E_2$ | $E_3$ | $E_4$ | $E_5$ | 收入 |
| ERACE | −0.132<br>(0.000) | −0.057<br>(0.001) | −0.231<br>(0.000) | −0.272<br>(0.000) | −0.333<br>(0.000) | 0.313<br>(0.000) |
| $E_1$ | | 0.563<br>(0.000) | 0.565<br>(0.000) | 0.662<br>(0.000) | 0.644<br>(0.000) | −0.270<br>(0.000) |
| $E_2$ | | | 0.444<br>(0.000) | 0.584<br>(0.000) | 0.560<br>(0.000) | −0.242<br>(0.000) |
| $E_3$ | | | | 0.586<br>(0.000) | 0.562<br>(0.000) | −0.354<br>(0.000) |
| $E_4$ | | | | | 0.660<br>(0.000) | −0.403<br>(0.000) |
| $E_5$ | | | | | | −0.420<br>(0.000) |
| 平均值<br>标准差 | −0.096<br>(0.544) | 0.052<br>(0.671) | 0.045<br>(0.666) | −0.111<br>(0.527) | −0.093<br>(0.547) | 6903.220<br>(4629.954) |

| | 对比编码的虚拟变量 | | | | | |
|---|---|---|---|---|---|---|
| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | 收入 |
| CRACE | 0.271<br>(0.000) | 0.132<br>(0.000) | 0.172<br>(0.000) | 0.164<br>(0.000) | 0.088<br>(0.000) | 0.313<br>(0.000) |
| $C_1$ | | 0.265<br>(0.000) | −0.257<br>(0.000) | −0.006<br>(0.358) | 0.027<br>(0.063) | 0.396<br>(0.000) |
| $C_2$ | | | −0.068<br>(0.000) | −0.002<br>(0.462) | 0.007<br>(0.343) | 0.270<br>(0.000) |
| $C_3$ | | | | 0.009<br>(0.310) | 0.068<br>(0.000) | 0.056<br>(0.001) |
| $C_4$ | | | | | 0.000<br>(0.491) | 0.105<br>(0.000) |
| $C_5$ | | | | | | 0.040<br>(0.012) |
| 平均值<br>标准差 | −0.021<br>(0.346) | 0.096<br>(0.544) | 0.150<br>(0.389) | 0.007<br>(0.706) | −0.018<br>(0.443) | 6903.220<br>(4629.954) |

注:括号里为相关系数的概率值。

人。效果编码虚拟变量的方差是两个待比较组群之间相对频数的函数，即 $p_j + p_{ref} - (p_j - p_{ref})^2$。那么，如上所述，$E_1$ 的方差是高级白领和初级白领之间相对频数的函数，即 $s^2_{E_1} = 0.1050 + 0.2006 - (-0.0956)^2 = 0.2965$，$s_{E_1} = 0.544$。

由 $E_1$ 到 $E_5$ 与收入的相关性系数可以看出，高级白领和体力劳动者之间的对比（即 $E_5$）是最强烈的，因为其均值偏离最大。但是，由于样本在各个职业和种族组群之间不是均匀分布的，所以零阶相关性的解释还不太明确。虚拟变量间的零阶相关性可以表示所对应组群的大小。在效果编码虚拟变量中，相关系数 0.50 只有在各个组群具有相同的样本量时才会出现。当参照组样本大于其组群时，零阶相关性系数会大于 0.50（例如，$E_1$ 和 $E_5$ 的或 $E_1$ 和 $E_4$ 的相关性）；当参照组样本量少于其他组群时，零阶相关性系数会降至 0.50 以下。

## 回归结果

虽然不同的编码方案会使回归系数在数值上有所不同，但是总体模型拟合度（由 $R^2$ 表示）以及种族和职业类别虚拟变量对收入的影响的显著性（由模型 1 的 $R^2$ 的 $F$ 检验和模型 3 相对模型 1 的 $R^2$ 增量指出）仍可重现第 3 章中模型 1 和模型 3 的结果。不同的编码方案会影响信息的捕捉方式，即对于不同编码方案，其组群间差异方式的排列不同，但是不会影响总体的结果，因为组群之间的结果差异相对于前面的估计保持不变，我们只是从另外一个角度来观察而已。

之前提到过，二进制编码虚拟变量只有在其他虚拟变量被控制的时候才能对参照组和指定组进行比较。换句话说，组群比较只能作为部分效果而存在。效果编码虚拟变量的情况是相似的，尽管比较的本质有所改变。这里，在控制了 $E_2$ 到 $E_5$ 后，$E_1$ 的偏回归系数提供了初级白领与样本中所有组群的比较结果。这样，由偏回归系数估计得出的数值就等于第 $j$ 组与没有加权的所有组群的收入期望值差异，即 $B_k = \overline{Y}_j - \sum \overline{Y}_j / j$，其中，$j$ 是原来名义度量的类别数，$\overline{Y}_j$ 是第 $j$ 类组群的均值。我们可通过截距得到没有加权过的所有组群均值的均值，该值可以作为所有组群差异的计算参考点。

没有加权的均值和总体样本的均值的度量不同，其数值相同与否，取决于组群均值相对于组群大小的变异性。总体的样本均值可以看成加权后的所有

组群均值的均值，因为我们计算样本均值时，是把每组的均值与该组的事件数相乘后求和，再除以总样本量所得出的。计算没有加权的均值的平均值，相当于给每一组样本赋予相同的权重 1，不管该组中的事件数为多少。这样做的结果是，一些只包含少量事件且度量不太准确的组群均值会与有大量观测值且度量较准确的组群均值得到同等对待，然而这种准确度上的差异会在回归系数的标准差中反映出来。该过程也可以使非标准化的回归系数独立于相应组群的大小。

**表 5.3 用不同编码方案的回归结果**

| | 效果编码 | | | 对比编码 | |
|---|---|---|---|---|---|
| | 模型 1 | 模型 3 | | 模型 2 | 模型 3 |
| 常数 | 6220.5<br>(85.8) | 6277.8<br>(83.2) | | 6567.7<br>(78.8) | 6751.5<br>(88.7) |
| 种族 | 1601.4<br>(85.8) | 838.0<br>(86.2) | | | 838.0<br>(86.2) |
| $E_1$ | | 853.5<br>(196.9) | $C_1$ | 5247.6<br>(223.4) | 6443.8<br>(260.3) |
| $E_2$ | | 129.2<br>(141.4) | $C_2$ | 1510.6<br>(137.2) | 2842.1<br>(271.1) |
| $E_3$ | | −908.9<br>(140.9) | $C_3$ | 1987.3<br>(192.4) | 1494.8<br>(196.3) |
| $E_4$ | | −1817.1<br>(211.4) | $C_4$ | 695.5<br>(102.1) | 519.1<br>(102.3) |
| $E_5$ | | −1952.2<br>(200.6) | $C_5$ | 172.2<br>(163.1) | 67.5<br>(161.1) |
| $R^2$ | 0.09792 | 0.24624 | | 0.22400 | 0.24624 |
| $F$ | 348.3 | 174.4 | | 185.0 | 174.4 |

注：括号里为回归系数的标准误。

现在，我们来考虑表 5.3 中模型 1 的回归结果。效果编码的种族变量(ERACE)是模型中唯一的自变量，所以，其截距等于黑人的收入均值加上白人的收入均值，然后除以 2。读者可以从表 2.2 的数字中证明这个数值。$B_{ERACE}$ = 1601.4 是白人收入期望值(7821 美元)和没有加权过的白人和黑人的平均值

的差异,即斜率。

模型3包含了ERACE和职业虚拟变量。从ERACE的回归系数可以看出,一旦考虑了不同职业类别的收入差异,白人的收入正效应就显而易见。同样,当控制了黑人/白人的收入差异时,初级白领和技术工人平均比其他组群的期望收入高,且初级白领的优势超过技术工人。操作工人、服务业工作者和劳工均在均值以下,其中服务业工作者和劳工的收入劣势最大。利用效果编码虚拟变量模型、收入和一个职业虚拟变量的偏回归系数,在控制其他职业虚拟变量时,可以被理解为该组的"反常"或者"独特"的性质度量(Cohen & Cohen, 1983)。通过对半偏相关系数取平方,我们可以评估每一个类别的特性使收入变异的程度。

如前几章所述,我们可以用每一特定组群的编码数乘以该组虚拟变量的估计系数来预测该组群的收入。如果虚拟变量是二进制编码的,该过程就可叙述为把虚拟变量的回归系数加入其所代表的组群,并丢弃所有被编码为0的虚拟变量。用效果编码虚拟变量,参照组统一被编码为-1,例如,高级白领在所有职业虚拟变量中就被编码为-1。因此,要计算白人高级白领的预测收入,我们有:

$$Y_{UWC} = 6277.8 + 838(1) + 853.5(-1) + 129.2(-1) - 908.9(-1) - 1817.1(-1) - 1952.2(-1) = 10811.3$$

同样,如果要计算黑人高级白领的期望收入,我们只需将ERACE(838)的系数乘以-1,其他不变,即可得出黑人预期收入为9135.3。通过对比表5.3中模型3和表3.2中模型3的期望收入值,读者可以证实,不论我们对虚拟变量使用哪种编码方案,所得出的期望收入值都是一样的。因此,效果编码虚拟变量和二进制编码虚拟变量的主要区别在于其参考点的定义。不同于估计每个组群相对某个特定参照组的差异,效果编码是把每个组群与该组中的所有组群进行比较。

先不考虑这些明显差异的解释,Suits(1983)证明了,通过在二进制编码虚拟变量的估计回归系数中加一个常数,就有可能变更解释框架,从而可以通过

各组群间没有加权过的均值解释所有组群的偏差。我们可以考虑最简单的情况，即用收入对一个二进制编码的虚拟变量种族(BLACK)进行回归后产生的结果，如表 3.1 模型 1 中所列出的：

$$Y_i = 7821.9 - 3202.9(\text{BLACK}) + e_i$$

通过一个常数 $c$ 对 $B_k$ 进行调整，我们可以把解释框架从二进制编码转化为效果编码模式。我们可以通过 $\sum(B_k + c) = 0$ 来确定 $c$ 的值，其中 $B_k$ 为二进制编码虚拟变量的回归系数，那么，$c = -(\sum B_k)/j$，其中 $j$ 为定性度量的类别个数。在该例中，$c$ 等于 $-(-3202.9/2) = 1601.45$。通过对每个虚拟变量的回归系数加一个 $c$，在常数项里减去一个 $c$，我们有：

$$Y_i = 6220.45 - 1601.45(\text{BLACK}) + 1601.45(\text{WHITE})$$

这样就可以表达 WHITE 的回归系数，尽管它在原先的模型设定中为参照组。我们可以简单地假设 WHITE 的效应在原先的规范中为 0。

当有多于一个的定性变量被加入规范中时，调整可以在虚拟变量组群内确定。比如，当用 INCOME 对 BLACK 和职业虚拟变量进行回归后，我们有：

$$Y_i = 10811.4 - 1676(\text{BLACK}) - 2842.1(\text{OCC}_2) - 3566.4(\text{OCC}_3)$$
$$- 4604.5(\text{OCC}_4) - 5512.7(\text{OCC}_5) - 5647.8(\text{OCC}_6) + e_i$$

$c_{\text{RACE}}$像之前那样确定，即 $c_{\text{RACE}} = -(-1676/2) = 838$。$c_{\text{OCC}} = -[(-2842.1) + (-3566.4) + (-4604.5) + (-5512.7) + (-5647.8)]/6 = 3695.58$。我们把 $c_{\text{RACE}}$加到种族的两个类别里，其中 WHITE 的回归系数根据原先的规范被设置成了 0，然后我们再把 $c_{\text{OCC}}$加到职业的 6 个类别里，其中高级白领的回归系数在原先的规范中也被设置成 0，之后我们从常数项中减去$c_{\text{RACE}}$和 $c_{\text{OCC}}$。这样，我们就可以表达任意组群相对于没有加权的所有组群的平均值的影响了。

$$Y_i = 6277.8 - 838(\text{BLACK}) + 838(\text{WHITE}) + 3695.6(\text{OCC}_1) + 853.5(\text{OCC}_2)$$
$$+ 129.18(\text{OCC}_3) - 908.9(\text{OCC}_4) - 1817.12(\text{OCC}_5) - 1952.2(\text{OCC}_6) + e_i$$

## 对比编码虚拟变量

表 5.1 下半部分表格列出了一系列对比编码虚拟变量。在 3 个条件下，研究者可以通过对比编码指定其感兴趣的比较：(1)对于含有 $j$ 类的名义变量，其表达需要指定 $j-1$ 个对比；(2)任何对比编码虚拟变量的编码的加和必须为 0；(3)任何两个虚拟变量的编码必须正交。根据经验产生的对比编码需要我们最初就将一系列类别归入两个合并的组群中。

对于此例，我们可以区分白领和蓝领工作者。$C_1$ 定义了所有白领与所有蓝领的比较。因为白领组合并了高级白领和初级白领，因此每个类别被编码为 0.5。同样，由于蓝领包含了其他 4 个组，所以每个类别被编码为 −0.25，负号表示蓝领与白领的对比。其中，0.25 是 4 个组群加入相等权重后产生的聚合群的结果，该 4 个编码之和为 1。

剩余的虚拟变量在它们最初的分类中定义了对比。比如，$C_2$ 对比了两个白领组的成员，因为每个组都是独立的，所以一组编码为 1，另一组为 −1。$C_3$ 比较的是技术工人和操作工人与服务业工作者和劳工之间的区别，前两组编码为 0.5，后两组编码为 −0.5。$C_4$ 定义了技术工人和操作工人之间的区别，$C_5$ 比较了服务业工作者和劳工之间的区别。[15] 我们可以将连续两对编码的乘积求和来检验该对比编码组的独立性。例如，对 $C_1$ 和 $C_2$ 的编码乘积求和，我们有 $(0.5)(1)+(0.5)(-1)+(-0.25)(0)+(-0.25)(0)+(-0.25)(0)+(-0.25)(0)=0$。

表 5.2 的下半部分表格列出了表 5.1 中定义的对比编码虚拟变量的零阶相关性、均值和标准差。其中，该对比编码虚拟变量的均值和标准差是组群相对大小的函数，但是由于编码规范中包括的数值均小于 1，因此这些频数和均值之间的关系对解释的用处不大。

如前一个例子，这一系列虚拟变量呈现出和其他变量的零阶相关性，即便之前它是用来定义组群对比而现在有了正交的性质。但是要求编码正交的条件与要求变量正交不同，对比编码虚拟变量之间的相关性是各组群相对大小的

函数。只有当观测值在各组中均匀分布时，相关性才为 0。[16]

有关对比编码虚拟变量和收入之间的零阶相关性的解释，确实不甚明确。如 $C_2$、$C_3$ 和 $C_4$，其解释基本上和效果编码虚拟变量一样，因为对于这 3 个变量，对比是由一个编码为−1 的组、一个编码为＋1 的组和其他编码为 0 的组而捕捉到的。这种编码方法和效果编码差不多，除了一点不同，即在对比编码虚拟变量中，−1 不是被分配到所有相同的组中的。因此，这些相关性是用来衡量平均收入编码为−1 和 1 的组间的差异程度，当然是在考虑了 INCOME 的方差的情况下。然而，如果组群大小不相等，编码为 0 的组也可以在其度量上影射该信息。

解释只有两种数值的编码的虚拟变量相关性（如 $C_1$）是非常直接的，因为 $C_1$ 把白领编码为 0.5，把蓝领编码成−0.25，收入和 $C_1$ 之间的零阶相关系数的平方测量了收入方差的比例，该比例可由白领和蓝领之间的差异来解释。当其他虚拟变量不稳定的时候，编码方案的目的是要计算加权后的各组群均值的均值。例如，由于高级白领和初级白领的编码是一样的，所以和零阶相关性有关的白领平均 INCOME 是一个合并的均值，忽略了原先细分的组群信息。换句话说，白领的均值是包括了所有白领工作者，即高级白领和初级白领加在一起的收入均值。因此，零阶相关性包括了加权后相似编码组的均值。我们知道，该数值可以通过每组的均值与每组的事件数相乘求和后，再除以所有事件数得出。然而，必须指出的是，该解释只有在处理仅有两个可能数值的对比编码虚拟变量的零阶相关性时才合适。所以，我们可以得出一个普遍的结论：尽管对比编码虚拟变量基于所有变量，为我们提供了总结回归结果的一个有用的替代方法，但是这种一次只能比较一个变量的简单描述性统计并不特别有用。

### 回归结果

表 5.3 的右边两栏列出了模型 2 和模型 3 的回归估计结果。没有包括模型 1 的估计结果，是因为它和效果编码虚拟变量完全一样。再提一下，第 3 章表 3.1 中列出的 $R^2$ 和 $F$ 检验可重现，这强调了 3 种方法在多元回归分析中的等价性。

包含在回归估计中的剔除过程相对二元测量，仍需要一个更直接的系数解释方法，尽管仍然需要一些额外计算。对模型截距的解释和对模型效果编码虚

拟变量的解释相同,即没有加权的所有组群均值的均值,它为我们提供了估计组群效应的参考点。每个虚拟变量指出了两个组群或组群集合的对比。变量的偏回归系数是没有加权的组群均值的均值与用来创建对比的编码差异的函数。由于剔除过程,编码为 0 的组已被排除,不进行比较。

组群对比 $C_j$ 被定义为:

$$C_j = B\frac{n_{g1} + n_{g2}}{(n_{g1})(n_{g2})} \qquad [5.1]$$

其中,$n_{g1}$ 为第一个集合中所包含的组群数,$n_{g2}$ 为第二个集合中所包含的组群数,$B$ 是虚拟变量的回归系数。例如,系数 $C_1$ 是白领和蓝领对比的函数,将其带入方程 5.1,我们得到:

$$C_1 = 5247.6\frac{2+4}{(2)(4)} = 5247.6(0.75) = 3935.7$$

其他对比的计算如下:

$$C_2 = 1510.6(2) = 3021.2$$

$$C_3 = 1987.3(1) = 1987.3$$

$$C_4 = 695.5(2) = 1391.0$$

$$C_5 = 172.2(2) = 344.4$$

通过表 2.2 列出的组群均值,读者可以证明,这些对比确实可以重现组群均值差异或者没有加权的组均值的均值。

这些对比的标准差可通过将回归系数的标准差乘以一个我们之前用于加权系数的因子而得到。例如,高级白领和初级白领的对比的标准差($C_2$)为 $(137.2)(2) = 274.4$。该值和表 3.1 中模型 2 列出的 $OCC_2$ 的数值一样。同时,回归系数的 $t$ 检验使我们可以估计由虚拟变量定义的对比是否可以推广到整体。在模型 2 中,回归系数 $C_1$ 到 $C_4$ 在 0.001 的水平上都很显著,但 $C_5$ 却不是。从而我们可以得出结论:总体而言,没有加权过的高级白领和初级白领的平均收入比蓝领的平均收入高;高级白领平均收入比初级白领高;没有加权的技术工人和操作工人组比服务业工作者和劳工组的平均收入高,但是劳工的期

望收入却与服务业工作者没有显著差别。[17]

当所有其他虚拟变量被控制时，半偏回归系数的平方表示了可由一个特定对比所解释的 $Y$ 的样本方差的比例。我们来看看表 5.3 中模型 2 的半偏系数，假设其最大部分的方差来自白领和蓝领之间的对比（例如，$0.365^2=13.3\%$）。然而，半偏相关系数的平方和不能为方程提供 $R^2$ 值，因为 $C$ 变量是相关的。只有当所有的组群大小都一样时，对比编码虚拟变量之间才是无关的，而且只有当回归因子正交时，半偏相关系数的平方和才等于 $R^2$。

# 第 6 章 | 虚拟变量用法专题

至此，我们一直在用同一个数据集的同系列变量来探索逐渐复杂化的模型中，虚拟变量的解释方法。在这一过程中我们也发现，引入虚拟变量可以使建模更加灵活。除了那些之前提到的假设，虚拟变量还经常用于一些其他形式。本章我们将探究一些虚拟变量在回归分析中的其他使用方法。

## logit 模型中的虚拟变量

越来越多的研究者在 logistic 回归模型中运用二分类或多分类的因变量。由于许多研究问题涉及组群差异，logistic 模型中的虚拟变量回归因子已很常见。假设一个模型以死亡率为因变量、性别为虚拟变量，那么 logit 回归的虚拟变量的回归系数代表一个和性别相关的死亡几率对数的增量或者减量。然而，解释一个几率对数并不像解释一个简单的机会比率那么有吸引力。那么，我们可以转换解释框架吗？答案是肯定的。我们可以通过 logit 系数的反对数来完成从相加效应（我们会在用线性相加模型预测几率对数时详细说明）到乘积效果，即把因变量变为一个简单的机会比率（Alba，1988）。由于对数转变可以使我们用相加的形式代表相乘关系，因此这种从相加关系到相乘关系的转变会伴随着从几率对数到机会比率的转变。

为了解释这个方法，我们先来看看 Idler 和 Kasl（1991）的研究结果，即上了年纪的女性的预期死亡率作为其自身健康和健康风险因素评估的函数。对于该模型，因变量为 4 年里的死亡机会比率。如果受访者去世了，则该比率编码为 1；若没有去世，则编码为 0。主观的健康状态由一组 3 个虚拟变量来衡量，以

“健康状况非常好”为参照组。健康状况控制了包括癌症、糖尿病、间歇性跛行和高血压，如果存在其中任一情况，则编码为1。此外，还有需要协助的活动个数、日常不可自理的活动个数(ADL)、身体质量指数、年龄以及测量现在或者以前是否吸烟的两个虚拟变量，其中，以不吸烟者作为参照组。表6.1记录了有关logistic回归结果的估计。右边栏为虚拟变量logit系数的反对数。我们可以看出，当控制了其他变量时，健康状况差的相对于健康状况非常好的，其死亡机会比率会增加。换句话说，在其他条件相等的情况下，健康状况差的女性的死亡几率是健康状况非常好的女性的3.12倍。同样，当控制了其他因素时，健康状况一般的女性的死亡几率是健康状况非常好的2.85倍，健康状态好的是健康状况非常好的2.37倍。此外，吸烟的净效应告诉我们，现在还在抽烟的女性的死亡几率是不抽烟女性的2.16倍。对于样本成员，过去吸烟然后戒掉的女性的死亡几率比不抽烟的女性低，即为不吸烟女性的0.73倍，然而估计差异并不显著。

**表6.1　虚拟变量的logistic回归**

| | logistic回归系数(*B*) | 反对数(*B*) |
|---|---|---|
| 常数 | −6.308*** | |
| 自我评估健康状况 | | |
| 差(同非常好比较) | 1.138* | 3.12 |
| 一般(同非常好比较) | 1.047* | 2.85 |
| 好(同非常好比较) | 0.862* | 2.37 |
| 糖尿病 | 0.963*** | 2.62 |
| ADL值 | 0.041 | |
| 活动 | 0.393* | |
| 间歇性跛行 | 0.982 | 2.67 |
| 高血压 | 0.369 | 1.45 |
| 年龄 | 0.061** | |
| 现在仍吸烟(同不吸烟比较) | 0.769*** | 2.16 |
| 以前吸烟(同不吸烟比较) | −0.312 | 0.73 |
| 体重(kg)/身高$^2$(m) | −0.076*** | |

注：* 表示系数在0.05显著性水平上显著；** 表示系数在0.01显著性水平上显著；*** 表示系数在0.001显著性水平上显著。
资料来源：Idler和Kasl(1991)。

模型中的因变量是进行了对数转变的，其虚拟变量系数解释要分两部分说明。第一部分（见第4章后半部分）通过系数的算术转变，读者可以解释其中的百分比差异。在这一章中，我们对乘积效应作出了解释。表面上来看，虚拟变量的系数在第4章运用的半对数模型中的解释和 logit 模型解释不同。相反，这两种解释只在方差上有微小的差别。在方程4.10中，我们把相对影响定义为存在某种特征的百分比差异，其由编码为1的虚拟变量说明。例如，在白人中，当控制了模型中的其他变量后，劳工的期望收入比高级白领的收入少40%。然而，如果不从回归系数的反对数里减1，我们的结论是针对职业差异的乘积项的，即劳工的收入是高级白领的60%。不管怎样，用虚拟变量回归系数估计的反对数转变模型来解释其相对影响是非常重要的。

## 曲线性检验

虚拟变量常用来代表名义编码的自变量类别。但是，我们也会用虚拟变量来表示定序变量或者区间变量。例如，当我们怀疑自变量与因变量之间存在非单调性或者曲线性关系，但又没有很好的基础来预测这种特殊形式的曲线性时，虚拟变量回归提供了一个非常有用的替代方法来取代多项式回归或者算术转变模型。通过用一系列虚拟变量代表一个定量自变量，我们把整体分布分成几小段，然后检验虚拟变量之间是否存在线性或者曲线性的关系。

现在，我们用一个比较熟悉的例子来解释收入和教育之间的关系。与其假设教育的影响在整个区域都是相同的，倒不如假设每增加一年的教育所带来的收入增量是由整个分布中，每增加一年教育年限所带来的影响而定的。为了检验这个教育和收入之间的曲线性关系，我们要评估两个模型：

模型6.1：$Y = f(\text{教育年限}) = \beta_0 + \beta_1 \mathrm{EDUC} + u_i$

模型6.2：$Y = f(\text{代表各教育水平的虚拟变量}) = \beta_0^* + \sum \beta_j ED_j$

在模型6.1中，教育（EDUC）是作为一个定量变量的，而在模型6.2中，教育被指定为一系列的虚拟变量，符号为 $\mathrm{ED}_j$。其中，当受访者没有接受过正式教育

时，$ED_0$ 编码为 1；当受访者完成了超过一年的教育时，$ED_1$ 编码为 1；当受访者已经完成了 17 年的正式教育时，$ED_{17}$编码为 1。参照组为接受了 18 年正式教育的受访者。

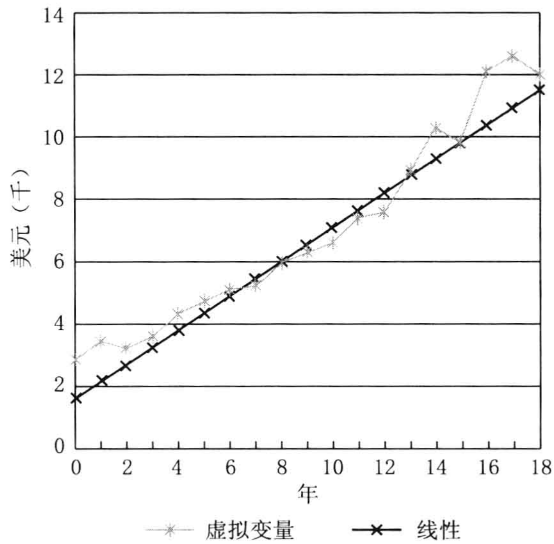



**图 6.1　非曲线性检验**

图 6.1 对这两个模型结果作出了解释。乍看之下，会感觉其关系确实有些曲线化。在整个教育区域内，期望收入差异在连续的组群之间是不一致的。斜率在低教育水平上比较平缓，随着教育程度的升高，其呈现出比较陡的趋势，尤其是完成了 12 年教育之后。用虚拟变量代表教育年限的模型相对于线性模型，可解释更多的方差。模型 6.1 解释了 19.6%的收入方差，而模型 6.2 解释了 21.4%的收入方差。当检验放宽了线性回归假设后，要检验被解释方差的增量是否统计上显著，就需要如下的 $F$ 检验，其中分子的自由度是由虚拟变量模型中多出的虚拟变量个数决定的，分母的自由度等于事件数减去虚拟变量模型中的参数数目：

$$F=\frac{(R_2^2-R_1^2)/(\mathrm{df}_2-\mathrm{df}_1)}{(1-R_2^2)/(N-\mathrm{df}_2)} \qquad [6.1]$$

将该 $F$ 检验运用到模型 6.1 和模型 6.2 中，我们得到：

$$F_{17\ 319\ 2} = \frac{(0.21390 - 0.19624)/(19-2)}{(1-0.21390)/3192} = \frac{0.00104}{0.00025} = 4.16$$

当在 0.01 显著性水平下与 $F$ 的临界值比较时，该 $F$ 值是统计上显著的。

如果自变量在度量上是真正连续的，原先的变量必须在用一系列虚拟变量表示前，重新编码成离散的类别。当这一步骤成为必要时，第二个模型和现在文中出现的模型或多或少会有些不同，因为原先的定量自变量和虚拟变量是一起包含在规范中的，但是 $F$ 检验的公式和两个方程的 $R^2$ 相比是不变的。

## 分段线性回归

虚拟变量可以让我们对一个回归线上斜率的突变建模。当斜率是逐渐变化的时候，即当 $Y_i$ 和 $X_{1i}$ 在 $X_{2i}$ 线性方式上增加或减少时，我们就可以用一个乘积项来捕捉该调解效应。当斜率突变时，我们可以用虚拟变量来协助估计该变化的强度和显著性。当我们可以界定出定量自变量($X_i$)分布的临界值，而且希望看到 $X_i$ 和 $Y_i$ 之间的关系在该临界值两边不同时，这个方法就相当有用。例如，零售业的营销人员经常可以从佣金中得到部分补偿，这些佣金与他们卖出的商品的数量有分级相关的关系。同样，投入和产出可能与经济规模有关，因此我们可根据该经济规模来修改某个产出水平下的投入方程。关于这个，我们会在后面具体说明。

假设我们有两个分布：第一个分布列出了全部产出；第二个分布列出了全部投入。我们可以进一步假设，当期望产量达到 5000 时，每单位投入会减少多少。因此，5000 即一个临界值 $X^*$ 。为了估计斜率，即每单位投入在 $X^* = 5000$ 时的变化，我们必须先通过临界值计算每个产出水平的偏差，即 $(X_i - X^*)$，然后定义一个虚拟变量($D_i$)，使其在产出超过临界值 5000 时为 1，否则为 0。那么，模型可以写成：

$$\text{模型 6.3：} Y_i = B_0 + B_1 X_i + B_2 (X_i - X^*) D_i + e_i$$

其中，$B_1$ 估计的是产出不到 5000 时的斜率，$(B_1 + B_2)$ 估计的是产出超过5000时的斜率。因此，$B_2$ 估计的是斜率的变化，$B_2$ 的 $t$ 检验提供给我们的是对斜率变化估计的显著性估测。

继续看我们的例子，当用投入对产出进行回归时，我们可以得到以下结果：

$$Y_i = \underset{(27.455)}{143.798} + \underset{(0.006)}{0.109}\text{OUTPUT} + e_i$$

该方程解释了 $Y_i$93.75%的方差，这说明，当产出增加一个单位值时，整体投入会增加大约 11 美分。换句话说，即每件物品的边际成本为 10.9 美分。当评估一个分段回归模型时，我们会发现：

$$Y_i = \underset{(34.264)}{87.059} + \underset{(0.010)}{0.129}\text{OUTPUT} - \underset{(0.018)}{0.045}(\text{OUTPUT} - \text{OUTPUT}^*)(D_i) + e_i$$

该方程解释了整体投入 95%的方差。和之前的模型相比，它有 1.25%的显著提高。另外，我们知道，当产出少于 5000 时，每单位的边际成本大约是 13 美分。然而，当产量高于 5000 时，边际成本会降到 8.4 美分(0.129−0.045)。

## 时间序列数据中的虚拟变量

当数据为截面数据时，虚拟变量可以提供一个在因变量期望值下估计组群差异的方法。在这种情况下，组群可由我们期望与因变量的分布结构相关的特征来定义。当数据成时间序列分布时，虚拟变量可以对组群观测进行分类。然而，在时间序列数据里的分组更有可能被定义为相对重要的事件。就像截面数据一样，虚拟变量如同分布机制的代理，不仅很难衡量，而且也很复杂。由于它们起着代理的作用，因此对差异背后的机制解释很容易引起争议，那么，其有效性及对任何模型设定的解释都可以成为一个反对的理由。

虚拟变量在时间序列回归中可用来捕捉区域性的或者组群差异的信息。但是，它们也可以用于检验参数结构稳定性和构建季节指数的工具。例如，研究美国工会会员增长的人员通常认为，《瓦格纳法案》的通过是工会主义发展的主

要原因;研究军方花费增长的调查人员发现,对战争动员的效应调整是非常必要的;还有一些建模研究试图说明,税法的变化是某项投资的盈利变化的决定性因素。我们知道,一个重要事件的影响可以导致趋势线的转变及其过程的结构调整。

关于结构稳定的例子,我们可以研究一下Carl Chen(1984)有关三里岛核事故对市场模型稳定性的影响问题。他的数据由1978年第一季度到1980年第一季度的70个公用事业股的周价格组成。这些核公司的股票在该事件后大幅度下降,Chen通过比较事件之前和之后的参数估计来检验市场模型的稳定性。根据需要,他提出了以下模型:

$$\text{模型 6.4:}\ r_{jt} = \beta_{j0} + \beta_{j1} r_{mt} + u_{jt}$$

其中,$r_{jt}$为$j$股在时间$t$内的每周回报,$r_{mt}$为以标准普尔指数表示的市场回报,$u_{jt}$是随机干扰项。

用虚拟变量来检验截距($\beta_{j0}$)和斜率($\beta_{j1}$)稳定性的方法需要我们把样本分成两个子样本期。在这种情况下,我们定义一个虚拟变量,当观测发生在事件之前时,$D=0$,发生在事件之后,则$D=1$,事件发生的那周从观测中删除。这样,检验模型变为:

$$\text{模型 6.5:}\ r_j = \beta_{j0}^{*} + \beta_{j1}^{*} r_m + \beta_{j2}^{*} D + \beta_{j3}^{*} r_m D + u_j$$

其中,$\beta_{j2}$估计了两个样本期截距值的差异,$\beta_{j3}$估计了两个样本期斜率系数的差异。研究核公司的组群,即那些在1980年有多余10%核燃料的结果如下,其中括号里的为$t$值:

$$\begin{aligned} r_j = &-0.0022 - 0.0031D + 0.3553r_m + 0.0614r_m D \\ &\ (1.43)\qquad (1.32)\qquad\quad (4.13)\qquad\ (0.50) \end{aligned}$$

根据这些结果,“没有结构性变化”的零假设不能被拒绝。

## 虚拟变量和自相关

我们可以考虑一个简单的时间序列模型,预测$Y_t$作为$X_t$和一个虚拟变量

$D$($D=1$ 表示后两个时期)的函数,该模型用来估计包含两个时期跨度的过程。那些对估计 $Y$ 水平的变化和形成 $Y$ 的过程变化感兴趣的人,可能会试图估计以下模型:

$$Y_t = B_0 + B_1 D + B_2 X_t + B_3 DX_t + e_t$$

其中,$B_1$ 估计了从第一期到第二期的变化,$B_2$ 估计了第一期时 $X$ 的效应,$B_3$ 估计了在第二期和第一期之间 ,$X$ 对 $Y$ 的影响变化的差异。

在估计时间序列模型时,研究人员必须注意它是否违反了没有自相关的假设。如果残差检验告诉我们误差之间有相关性,则 OLS 估计是无效的。要处理自相关问题,研究者必须经常假设干扰项是从一阶自相关过程中产生的,也就是说,现时段干扰项是之前时段干扰项的函数。这种相关的程度由自相关系数 $\rho$ 来测量。那么,补救措施就会涉及估算广义差分方程,即用($Y_t - \rho Y_{t-1}$)对($X_t - \rho X_{t-1}$)进行回归。其中,$\rho$ 为对自相关系数的估计。但是虚拟变量呢?是不是要对它们进行同样的转变呢?其实不用(参见 Maddala, 1992:321—322)。假定虚拟变量定义了两组观测,最关键的观测为第二期中的第一个观测,研究人员必须对这些组群里的观测进行如下定义:

第一,对于所有第一期的观测,$D$ 值为 0;第二期的第一个观测的 $D$ 值为 $1/(1-\rho)$;其他第二期的观测的 $D$ 值为 1。

第二,对于所有第一期的观测的乘积项,$DX_t$ 的值为 0;第二期第一个观测的乘积项为 $X_t$;对于其他第二期的观测,其乘积项值为($X_t - \rho X_{t-1}$)。

# 第7章 | 结论

我们知道,恰当地运用虚拟变量可以大大提高回归模型的灵活性。然而,恰当地使用和解释虚拟变量涉及很多复杂的问题。本书的目的是通过分析逐步复杂的情况,为读者提供一些使用虚拟变量的指导。这些方法当然不完全,因为我们仅把视线局限于虚拟变量作为单方程模型的自变量中。但是,因子分析中的二元变量使用方法,如结构方程体系中的内生和外生变量,或者单方程或多方程系统中的因变量受到越来越多的关注。想追求更高深的定量数据分析的读者可以参考如下文献:Maddala(1983)处理回归模型中多线因变量的方法;Haberman(1978、1979)的两卷书;Goodman(1978)有关定量数据分析的文章;Aldrich 和 Nelson(1984)关于线性概率、logit 和普罗比模型概论;Allison(1984)有关事件历史模型的讨论;Muthen(1984)和 Shockey(1988)对无法观测变量的离散数据模型的讨论;Clogg 和 Goodman(1984、1985)的潜在结构分析以及 Winship 和 Mare(1983、1984)关于离散数据的结构方程模型和回归模型的文章。

# 注释

[1] 为了产生独立可靠的估计,就必须保证有足够数量的黑人。因此,我们对黑人区的住户进行过度采样。简单来说,在该测试中,我们用的是没有进行加权的数据。

[2] 实际意义较次要的是剩余的职业类别虚拟变量的相关系数。因为它们代表了所有不同且互斥的单一属性类别(如职业),这些虚拟变量都必须是逆向联系的,即相关性为负。在二分变量中,如同该例的种族,BLACK 和 WHITE 之间的相关性为$-1.00$。对于多分类变量,尽管相关性仍为负,但却不可能负得如此完美。如果一个受访者为服务业工作者,那他一定不是初级白领、技术工人、操作工人或者劳工,但是如果一个受访者不是服务业工作者,那他不一定是操作工人、劳工等等。两个虚拟变量之间的相关性大小是每两个变量中被编码为 1 的事件数和样本大小的函数。在表 2.3 中,技术工人和操作工人包含的事件最多,因此这两类的相关性最高,为$-0.328$。相反,服务业工作者和劳工是事件数最少的两类,其相关性仅为$-0.108$。在这种情况下,两个虚拟变量的相关性等于$-[(p_j p'_j)/(1-p_j)(1-p'_j)]^{1/2}$。

[3] 我们必须记住,相关系数对变量的方差非常敏感。对于虚拟变量,相关系数由类别的相对频数而定。

[4] 由于模型 1 为一个二元回归模型,因此,$F$ 检验和 $t$ 检验是等价的,该检验的 $t$ 值($-18.7$)是 $F$ 值(348.3)的平方根。从模型回归估计得到的信息和从单因素方差分析得到的结果一样,当相同的组均值估计出现时,$F$ 检验会得出相同的结果(同样的数值)。另外,$\eta^2$(在该例中等于 0.09792)和模型 1 中的 $R^2$ 也是一样的。

[5] $F$ 检验的自由度包括了与回归平方和及残差平方和有关的自由度。回归平方和的自由度和模型中自变量的数目是相等的,在该情况下,如果模型包含 5 个虚拟变量,那么自由度为 5。残差平方和的自由度等于 $N-k-1$,其中 $N$ 是观测数,$k$ 是模型中自变量的个数。当 $F$ 检验是由 $R^2$ 和 $(1-R^2)$ 的比率计算出来的时候,自由度则如上所述。

[6] 从数学的角度来讲,参照组的选择是随意的,研究者可以选择不同的参照组再进行回归,让计算机程序来提供利益 $t$ 检验。

[7] 方程 4.4 通过两个系数的加和来捕捉黑人在不同职业类别中的期望收入差异。

[8] 事实上,黑人中由显著职业差异导致的平均收入差异很少,该现象大部分归因于另外两个被控制了的变量——教育年限和工作任期。如果我们所估计的模型只包含种族、职业和种族与职业的乘积项,那么回归结果应该为:

$$
\begin{aligned}
E(Y_i) = {} & 10960.3 - 3958.4(\mathrm{BLACK}) - 2898.9(\mathrm{OCC}_2) - 3625.6(\mathrm{OCC}_3) - 4875.0(\mathrm{OCC}_4) \\
& - 6154.7(\mathrm{OCC}_5) - 6182.9(\mathrm{OCC}_6) + 1747.8(\mathrm{BLOCC}_2) + 1781.6(\mathrm{BLOCC}_3) \\
& + 2594.5(\mathrm{BLOCC}_4) + 3238.6(\mathrm{BLOCC}_5) + 2885.4(\mathrm{BLOCC}_6)
\end{aligned}
$$

其所有系数的 $t$ 值都比 $\pm 2.00$ 大。用方程 4.5 来构造黑人在不同职业中平均收入差异相当的 $t$ 检验,我们发现,只有初级白领和高级白领的差异没有达到 0.05 的显著水平。读者也可以用以上估值来证明该模型设定中的均值和由种族内不同职业计算出来的均值(见表 2.2)是等价的。

[9] 教育年限和工作的差异效应与职业类别相关的问题是可能的。在该例中,我们对黑人和白人之间的差异估计感兴趣,不仅在于其期望收入的差异,还有和收入水平相关的结构影响问题。除了看种族之间的教育效应,我们还需在假设中加入职业类别的影响。在同一个雇主下多工作一年,其作为稳定性指标、作为对公司的忠诚或者特定的职业训练,可能技术工人相对于劳工会得到更多的收入。多一年的教育年限,专业人士可以挣得更高的收入,但这一点却不适用于工厂操作工

人，因为这些不会增加他们职位提升的机会。如果我们的兴趣在于教育和工作任期的职业影响，乘积项可以根据每个职业虚拟变量和教育年限编码，也可以对每个职业虚拟变量和工作任期的乘积项进行编码。如果我们只检验职业差异效应并且发现教育和工作期限的差异影响非常显著，那么，就要构建 6 对平行线来看教育差异效应，每对实线都针对一个特定的职业组群，其斜率对于黑人和白人是一样的，但是对于不同的职业类别，斜率是不同的。关于描述工作任期差异效应的虚线，其趋势也一样。

[10] 反对一个零假设，即 $\beta=0$，对 $-0.16$ 进行 $t$ 检验会产生一个 $-1.45$ 的 $t$ 值；对 $-0.32$ 进行 $t$ 检验，会产生一个 $-2.29$ 的 $t$ 值。

[11] 有关两个种群参数的等效检验已经提出。其中一个检验适应于方差检验的框架(Chow，1960；Maddala，1992)。其步骤是我们先分别估计每个组群的回归模型，然后从不同的回归结果中获得其残差平方和($RSS_j$)。并且，还需要对合并的样本进行回归估计，同时得到合并样本的残差平方和 RSS。对于该参数等效的 $F$ 检验(有 $k+1$ 和 $n_1+n_2-2$ 个自由度)，$F=[(RSS_{pooled}-\sum RSS_j)/(k+1)]/[\sum RSS_j/(n_1+n_2-2k-2)]$，其中，$\sum RSS_j$ 为对不同组群进行回归估计得到的 RSS 之和，$k$ 为模型中自变量的个数，$n_1$ 和 $n_2$ 分别为两个组群的观测数。该例的 $F$ 值已经大到可以拒绝零假设(参数的平等性)，这说明在两组群中，不是所有的自变量的影响都一样。然而，该检验并不能表明哪些参数是不同的。

[12] 残差平方和 RSS/$(n-k-1)$ 中，$n$ 为事件数，$k$ 为模型中的自变量个数，RSS 是残差平方和，可由 $\sum e_i^2$ 算得，但是一般的统计软件程序都会提供该结果。在 OLS 假设下，该数通常可以提供对 $\sigma^2$ 的无偏估计，即总体的方差 $u_i$。

[13] 在计算对数时，我们必须说明其基数。最普遍的基数就是 e，其通常也被称为“自然对数”，还有一个就是 10。e 的值约为 2.72。要对以 e 为底的 $X$ 取对数，我们需要确定 $X$ 所需要的 e 的幂次。同样，如果取以 10 为底的 $X$ 的对数，我们就需要确定 $X$ 所需要的 10 的幂次是多少。对数模式会使分布变得与之前不同，其非线性也非常明显，因为以 10 为底的对数值 1、2、3 对所对应的最初分布数值为 10、100 和 1000。

[14] 估计黑人在不同职业内影响的显著性与确定黑人和白人在不同职业内差异的显著性的过程是一样的。黑人在高级白领中的效应可以通过 BLACK 回归系数捕捉到，在初级白领中的效应可以看 $B_1+B_9$；对于技术工人，该效应为 $B_1+B_{11}$；对于服务业工作者，为 $B_1+B_{12}$；对于劳工，黑人的效应为 $B_1+B_{13}$；如果要看种族效应在职业类别内是否显著，即看其相对其他特定效应的净效应，我们要用方程 4.5 的 $t$ 检验来表达。5 个职业类别的 $t$ 值分别为 $-4.07$、$-5.58$、$-5.63$、$-1.90$和$-4.70$。

[15] 当创建对比编码变量时，我们有很多选择。比如，可以用技术工人与操作工人、服务业工作者和劳工的组合来进行比较。这样，技术工人则被编码为 1，而操作工人、服务业工作者和劳工就被编码为$-1/3$。之后，再比较操作工人与服务业工作者和劳工的组合，这样，最后一个比较也和文中所述的一样了。

[16] 熟悉传统方差检验的读者可能知道，这种情况和同组群大小要求的情况的相似处在于，它们都是在一个 $n$ 因素方差分析中作出正交设计的。

[17] 由 $C_5$ 定义的对比显著性的替换性检验，在第 3 章描述关于服务业工作者和劳工之间的差异时已被提及。方程 3.1 为包含二进制编码虚拟变量的两组群间的差异提供了检验方程。读者可以用表 3.1 中的结果来证实两个过程的 $t$ 值相同，而且其对比和标准差都与方程 3.1 中定义的均值差异和标准差差异估计一样。

# 参考文献


Alba, R. D. (1988) "Interpreting the parameters of log-linear models," in J. S. Long (ed.) *Common Problems/Proper Solutions: Avoiding Error in Quantitative Research*. Newbury Park, CA: Sage.

Aldrich, J. H., and Nelson, F. D. (1984) *Linear Probability, Logit, and Probit Models*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—045. Beverly Hills, CA: Sage.

Allison, P. D. (1984) *Event History Analysis: Regression for Longitudinal Event Data*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—046. Beverly Hills, CA: Sage.

Amemiya, T. (1986) *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Bartlett, M. S. (1937) "Properties of sufficiency and statistical tests." *Proceedings of the Royal Society of London 160* (Series A).

Berry, W. D., and Feldman, S. (1985) *Multiple Regression in Practice*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—050. Beverly Hills, CA: Sage.

Bohrnstedt, G., and Knoke, D. (1982) *Statistics for Social Data Analysis*. Itasca, IL: F. E. Peacock.

Chen, C. (1984) "The structural stability of the market model after the Three Mile Island accident." *Journal of Economics and Business* 36:133—140.

Chow, G. C. (1960) "Test of equality between sets of coefficients in two linear regressions." *Econometrica* 28:591—605.

Clogg, C. C., and Goodman, L. A. (1984) "Latent Structure analysis of a set of multidimensional contingency tables." *Journal of the American Statistical Association* 79:762—771.

Clogg, C. C., and Goodman, L. A. (1985) "Simultaneous latent structure analysis in several groups," in N. Tuma (ed.) *Sociological Methodology*. San Francisco: Jossey-Bass.

Cohen, J., and Cohen, P. (1983) *Applied Multiple Regression* (2nd *ed.*). Hillsdale, NJ: Lawrence Erlbaum.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981) "A comparative study of tests for homogeneity of variances with applications to the Outer Continental Shelf bidding data." *Technometrics* 23:351—361.

Darlington, R. B. (1990) *Regression and Linear Models*. New York: McGraw-Hill.

Dunn, O. J. (1961) "Multiple comparisons among means." *Journal of the American Statistical Association* 56:52—64.

Glejser, H. (1969) "A new test for homoscedasticity." *Journal of the American Statistical Association* 64:316—323.

Goldfeld, S. M., and Quandt, R. E. (1972) *Nonlinear Methods in Econometrics*. Amsterdam: North-Holland.

Goldfeld, S. M., and Quandt, R. E. (1978) "Asymptotic tests for the constancy of regressions in

the heteroscedastic case." *Research Memorandum* No. 229, Econometric Research Program, Princeton University.

Goodman, L. A. (1978) *Analyzing Qualitative/Categorical Data*. Cambridge, MA: Abt Associates.

Gujarati, D. N. (1970) "Use of dummy variables in testing for equality of sets of coefficients in two linear regressions: A note." *American Statistician* (February).

Gujarati, D. N. (1988) *Basic Ecnomiteirsc* ($2^{nd}$ *ed.*). New York: McGraw-Hill.

Haberman, S. J. (1978) *Analysis of Qualitative Data*, Vol. 1: Introductory Topics. New York: Academic Press.

Haberman, S. J. (1979) *Analysis of Qualitative Data*, Vol. 2. New York: Academic Press.

Halvorsen, R., and Palmquist, R. (1980) "Interpretation of dummy variables in semilogarithmic equations." *American Economic Review* 70:474—475.

Idler, E. L., and Kasl, S. (1991) "Health perceptions and survival: Do global evaluations of health status realy predict mortality?" *Journal of Gerontology* 46(2):S55—65.

Jaccard, J., Turrisi, R., and Wan, C. K. (1990) *Interaction Effects in Multiple Regression*. Sage University Paper series on Quantitative Applicatons in the Social Sciences, 07—072. Newbury Park, CA: Sage.

Johnston, J. (1984) *Econometric Methods* ($3^{rd}$ *ed.*). New York: McGraw-Hill.

Kendall, M. G., and Stuart, A. (1979) *The Advanced Theory of Statistics*, Vol. 2($4^{th}$ ed.). New York: Charles Griffin.

Kmenta, J. (1986) *Elements of Ecnonomtrics* ($2^{nd}$ *ed.*). New York: Macmillan.

Levene, H. (1960) "Robust tests for equality of variances," in I. Olkin(ed.) *Contributions to Probability and Statistics*. Stanford, CA: Stanford University Press.

Lewis-Beck, M. S. (1980) *Applied Regression: An Introduction*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—022. Beverly Hills, CA: Sage.

Long, J. S., and Miethe, T. D. (1988) "The statistical comparison of groups," in J. S. Long (ed.) *Common Problems/Proper Solutions: Avoiding Error in Quantitative Research*. Newbury Park, CA: Sage.

Maddala, G. S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Maddala, G. S. (1992) *Introduction to Econometrics* ($2^{nd}$ *ed.*). New York: Macmillan.

Miller, R. G., Jr. (1966) *Simultaneous Statistical Inference*. New York: McGraw-Hill.

Muthen, B. (1984) "A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators." *Psychometrics* 46:115—132.

Ryan, T. A. (1960) "Significance tests for multiple comparisons of proportions, variances, and other statistics." *Psychological Bulletin* 57:318—328.

Schroeder, L. D., Sjoquist, P. L., and Stephan, P. E. (1986) *Understanding Regresion Analysis: An Introductory Guide*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—057. Beverly Hills, CA: Sage.

Shockey, J. W. (1988) "Latent-class analysis: An introduction to discrete data models with unob-

served variables," in J. S. Long(ed.) *Common Problems/Proper Solutions Avoiding Error in Quantitative Research*. Newbury Park, CA: Sage.

Suits, D. (1983) "Dummy variables: Mechanics v. interpretation." *Review of Ecnomics and Statistics* 66:177—180.

Welch, B. L. (1938) "The significance of the difference between two means when the poplation variance are unequal." *Biometrika* 29:350—362.

Winship, C., and Mare, R. D. (1983) "Strutural equations and path analysis for discrete data." *American Journal of Sociology* 89:54—110.

Winship, C., and Mare, R. D. (1984) "Regression models with ordinal variables." *American Sociological Review* 49:512—525.

## 译名对照表

| | |
|---|---|
| additive effect | 相加效应 |
| bias | 偏差 |
| binary coding | 二进制编码 |
| bivariate regression | 二元回归 |
| chance fluctuation | 随机波动 |
| contrast coding | 对比编码 |
| correlation | 相关性 |
| covariance | 协方差 |
| cross sectional data | 截面数据 |
| cross-tabular analysis | 列联表分析 |
| curvilinearity | 曲线性 |
| dependent variable | 因变量 |
| descriptive Statistics | 描述性统计 |
| differential effects | 差异效应 |
| discrete variable | 离散变量 |
| dummy variable | 虚拟变量 |
| effects coding | 效果编码 |
| expected value | 期望值 |
| explanatory power | 解释功效 |
| Fisher's protected $t$ method | Fisher 的 $t$ 保护方法 |
| fit of the model | 模型拟合 |
| homogeneity | 同质性 |
| hypothetical data | 假设数据 |
| independent variable | 自变量 |
| incremental $F$ test | 增量 $F$ 检验 |
| interaction | 交互作用 |
| interaction effect | 交互效应 |
| interaction term | 交互项 |
| intercept | 截距 |
| magnitude | 强度 |
| marginal value | 边际值 |
| mean regression sums of squares | 平均回归平方和 |
| measures of association | 相联度量 |
| midrange mean value | 中距均值 |
| moderating effect | 调解效应 |
| multiplicative effect | 乘积效果 |

| | |
|---|---|
| nominal variable | 名义变量 |
| non-independent test | 非独立检验 |
| non—normality | 非常态性 |
| null hypothesis | 零假设 |
| one-way analysis of variance | 单因素方差分析 |
| orthogonal | 正交 |
| partial correlation | 偏相关 |
| partial effect | 局部效应/偏效应 |
| partial slope | 偏斜率 |
| partialing procedure | 剔除过程 |
| piecewise linear regression | 分段线性回归 |
| point biserial correlation coefficient | 点二列相关系数 |
| point estimate | 点估计 |
| polynomial regression | 多项式回归 |
| polytomous variable | 多分类变量 |
| population | 整体 |
| pooled estimate | 合并估计值 |
| power | 检定力 |
| reference group | 参照组 |
| regressor | 回归因子 |
| regression coefficient | 回归系数 |
| residual sum of square | 残差平方和 |
| robustness | 稳健性 |
| sampling error | 抽样误差 |
| semilogarithmic equation | 半对数方程 |
| simultaneous statistical inference | 同时统计学推论 |
| single explanatory characteristic | 单解释性特征 |
| slope | 斜率 |
| statistically significant | 显著性水平 |
| *t* statistic | *t* 统计量 |
| test statistic | 检验统计量 |
| time-series data | 时间序列数据 |
| two-tailed test | 双尾检验 |
| variance | 方差 |
| variation | 变异 |
| weight | 加权 |

# 多元因果中的交互作用

## 作者简介

### 詹姆斯·杰卡德(James Jaccard)

美国纽约州立大学阿尔巴尼分校心理学特聘教授。主要研究领域是种群动态心理学、青少年危险行为、青少年意外怀孕和青少年醉酒驾驶,其研究关注点是基于家庭视角来处理青少年的问题行为。

### 罗伯特·图里西(Robert Turrisi)

美国博伊西州立大学心理学教授、家庭研究项目主任。其研究焦点是把心理学的行为决策理论应用到青少年的健康问题和家庭关系上,包括发展和评估针对大学生酗酒、青少年醉酒驾驶和以父母为基础的干预的有效性。

## 译者简介

### 蒋　勤

曾先后获中国青年政治学院法学学士学位、北京大学社会学硕士学位,现为香港科技大学社会科学部博士候选人。

# 序

经典回归模型的表达式很直接，以最常见的三变量多元回归为例，$Y$ 是因变量，$X_1$ 和 $X_2$ 是自变量，可把方程 1 写成如下形式：

$$Y = a + b_1X_1 + b_2X_2 + e \qquad [1]$$

这里的 $a$、$b_1$ 和 $b_2$ 都是待估计的总体参数，$e$ 是残差。假设满足必要的假定，那么普通最小二乘法(OLS)估计就是最佳的线性无偏估计。这个 OLS 方程是加叠性的，各项共同解释 $Y$，没有自变量相乘的项。假设自变量相乘产生乘积项 $X_1X_2$，那么就可改写上面的方程为方程 2：

$$Y = a + b_1X_1 + b_2X_2 + b_3(X_1X_2) + e \qquad [2]$$

现在，我们有了一个既包括加叠性变量，又包括非加叠变量的方程。而乘积项或乘法的项目($X_1X_2$)也被称为“交互作用项”，它的系数 $b_3$ 估计交互作用。问题是，它的含义是什么呢？

通常，社会科学专业一年级的研究生会觉得这个问题挺难回答。他们这么理解加叠模型(方程 1)：当 $X_2$ 保持不变时，$X_1$ 变化一个单位，会导致 $Y$ 变化 $b_1$ 个单位。但如何理解 $b_3$，即方程 2 中非加叠项的系数？当 $X_1$ 对 $Y$ 的影响取决于 $X_2$ 的取值时，就存在交互作用。$b_3$ 估计的正是这种交互作用。请考虑这个例子：$Y$ = 个人对竞选运动的捐献额(美元)，$X_1$ = 收入(美元)，$X_2$ = 教育程度(0 = 未上大学，1 = 大学及以上)。如果研究者认为对受过大学教育的人而言，其收入对其捐献额影响更大(相比未受过大学教育者收入的作用)，那么研究者会偏向于使用包含交互作用的模型设定(方程 2)，而非方程 1。

很明显，交互作用假设——一个变量的影响取决于另一个变量的取值——应当需要检验。回顾已发表的研究，我们发现事实并非如此(很多研究不检验变量间的交互作用)，而我认为其原因在于交互作用这个想法较难捕捉，因此本书非常有意义。Jaccard 博士和 Turrisi 博士使理解和估计回归模型中的交互作用变得非常清楚。比如，他们强调，仅仅通过调节变量在不同组之间的作用差异来检验交互作用是不够的，如在上例中，仅仅分别观察上大学和没上大学两组人的收入在(分别的)回归方程中对政治捐献的影响是不够的。进而，他们提供了一个常常被忽略的见解：真正的度量来自数据本身，而不是测度。

因此，要用回归检验交互作用，重要的是测度接近数据内在定距属性的程度。同时，本书也有许多新主题，比如聚类数据的交互作用以及随机系数模型。事实上，仔细阅读参考文献部分，我们能发现，1990 年之后出版的文献有 30 多种。

研究实践者必须利用工具去假定、理解和估计交互作用。仔细学习此书能使研究者在回归分析中达到上述要求。若要在其他或更高级的研究情境下研究交互作用，读者可参考 Jaccard 教授的其他著作。

迈克尔·*S*. 刘易斯-贝克

# 第 1 章 | 导论

许多社会科学理论框架都关注因果模型。这些模型试图设定一个或多个自变量对一个或多个因变量或结果变量的影响。简单地说，因果模型存在 6 种可能关系，如图 1.1 所示。直接因果是一个变量 $X$ 为另一个变量 $Y$ 的直接原因。在理论系统语境下，$X$ 就是 $Y$ 的直接决定因素。间接因果是 $X$ 对 $Y$ 有因果作用，但 $X$ 对 $Y$ 产生的作用只能通过另一个变量 $Z$ 来实现；虚假因果就是 $X$ 和 $Y$ 相关，但这种相关性完全是因为存在一个共同原因 $Z$，即 $X$ 和 $Y$ 不存在真正的因果关系；互为因果是指 $X$ 对 $Y$ 有因果影响，$Y$ 对 $X$ 也有因果影响；待定因果是指 $X$ 和 $Y$ 相关，但这种关联的来源并未得到确认；最后，调节因果指 $X$ 和 $Y$ 之间的关系受到另外一个变量 $Z$ 的调节作用。换句话说，$X$ 和 $Y$ 的关系是变动的，这种变动取决于 $Z$ 的取值。

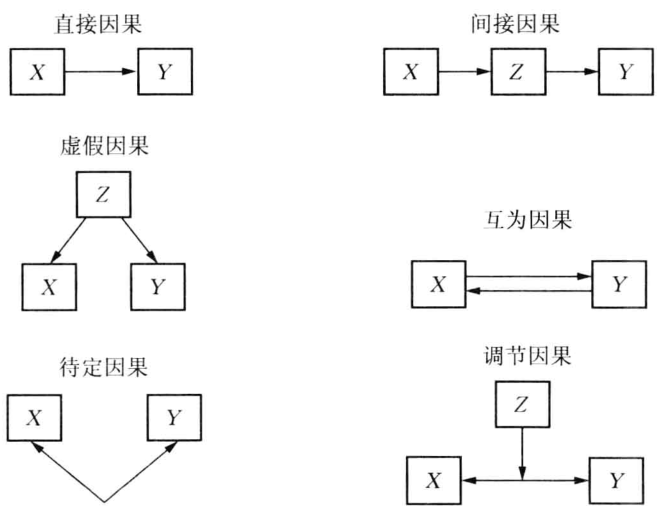



**图 1.1　因果关系的例子**

本书考察的是调节关联的统计分析，关注至少有一个自变量（或解释变量）是连续变量的情况。调节关系常被称为“交互作用”，尽管不同统计模型对交互作用的精确概念化有所不同。我们的关注点在于分析多元回归中的调节关联。在使用多元回归方法进行涉及连续变量的调节关联分析时，存在一些困惑需要厘清。

现有文献充斥着大量关于检验这些模型最佳方案的、相互矛盾的建议和忠告。本书最主要的目的就是梳理有关此主题的零散文献，说明进行包含连续变量调节关系的分析所涉及的中心议题。我们的目标是为那些具备多元回归操作知识的社会科学研究者展现一个合理的、易操作的导论。

在这一章中，我们会简单地回顾社会科学研究中交互作用分析的关键概念。首先考察一般意义上的交互作用概念，然后分析简单主效应和交互对比的概念。我们将回顾基本的多元回归程序，为下面各章节提供一个框架，包括虚拟变量的使用、简单变换对回归系数的影响以及预测值的计算。

## 交互的概念

如上面指出的，在社会科学中，许多方法都对交互作用进行了概念化，并且在哪种方法是最佳方法这一点上，存在很大的争议。一个流行的概念化交互作用的流派是把它当做调节关系。这一视角可用一个三变量系统进行展示，在此系统中，第一个变量被视为结果变量，第二个变量被视为自变量，第三个变量被视为调节变量，并且，结果变量被认为受到自变量的影响。而当自变量对因变量的影响取决于第三个变量的取值时（这一变量又称调节变量），就存在交互作用。例如，受教育年限对收入的影响可能因民族的不同而不同。教育对某些民族收入的影响要比另一些民族大。在调节变量框架中，收入是结果变量，受教育年限是自变量，而民族就是调节变量。再如，社会阶级对人们使用保健站频率的影响可能因性别的不同而不同。在此个案中，某人使用保健站的频率是结果变量或者因变量，社会阶级是自变量，而性别则是调节变量。性别被认为能够“调节”社会阶级对保健站使用情况的影响。

利用调节变量的方法进行交互分析要求理论家确定一个调节变量以及所谓的关键自变量。关键自变量对因变量的影响被认为根据调节自变量的某种函数关系而变化。大部分正式研究都很容易确定一个解释变量作为"调节者"。例如,研究者想要确定情绪低落的临床治疗是否对男性比对女性更有效。在此个案中,很明显性别是调节变量,而实施治疗与否是关键自变量。

当一个理论家的调节变量是另一个理论家的关键自变量(或反之)时,那么就出现问题了。例如,一个研究产品质量与产品选择的消费心理学家或许对产品质量如何影响产品购买决定,而这又如何受到产品价格的调节作用感兴趣。相反,一个市场研究者使用消费心理学家的实验范式,但他可能对产品价格影响产品购买行为,而这种影响又如何受到产品质量的调节作用更感兴趣。在这两个例子中,研究者只是依据实践和理论方面的指导来确定调节变量,并没有设定一个比另一个好。而且,从统计上讲,这两种概念化方式下的交互分析会得到同样的统计结果。这两种设定方式代表的是针对同一现象的不同视角,不同视角引导研究者强调数据的不同方面。

有时,理论家也不确定哪个变量应当具有调节者的地位。假设研究者检验的是性别和民族对堕胎态度的影响,他/她想要确定的是对堕胎态度的性别差异如何作为民族的一个函数而变动。在此情况下,性别是关键自变量,民族是调节变量。研究者或许又会对堕胎态度的民族差异作为性别的一个函数感兴趣,在此情况下,民族是关键自变量,性别是调节变量。没什么能阻止研究者在这种情况下使用任一方案来刻画数据,这只不过涉及我们到底对什么感兴趣的问题。在第 4 章中,还会深入讨论此问题。

Jaccard 和 Dodge 讨论了其他概念化统计交互作用的方法。调节方法在大量的研究领域被援引,甚至研究者有时并未意识到他们在这么做。在应用性研究情境下给交互作用灌输实质含义有些困难,大部分成功者最终都回到了调节变量的框架下。既然这一框架如此流行且易于理解,本书也将采用这一框架。第 4 章回顾了其他思考交互作用的框架。

## 简单效应与交互比较

在调节变量分析中有两个重要概念,简单效应(亦称"简单主效应")和单自由度交互比较。展示这些概念的最好途径是使用定性解释变量的析因设计。假设某社会科学家找到 300 名已婚人士,一半为男性,一半为女性(但男性与女性均为非配偶关系)。

表 1.1 性别与宗教对堕胎态度的影响

| | 天主教 | 新 教 | 犹太教 |
|---|---|---|---|
| 女 性 | 5.0 | 6.0 | 7.0 |
| 男 性 | 3.0 | 3.0 | 3.0 |

在这些人中,1/3 是天主教徒,1/3 是新教徒,1/3 是犹太教徒。这形成了一个 2×3 的由性别和宗教交叉而成的析因设计。令研究者感兴趣的是性别和宗教如何与堕胎态度联系在一起,堕胎态度用 0 到 10 的尺度来测量,分数越高代表越赞同。每个子群体的均值列在表 1.1 中。假设在这个特殊的例子中,研究者决定把宗教概念作为调节变量,而把性别作为关键自变量。

### 简单效应

研究者利用这些数据可以处理很多问题。一个常见问题是,关键自变量(性别)是否在调节变量的每一个类别上都有不同的影响?要回答这个问题,就需要单独比较天主教男性和女性的均值以及新教徒的均值,最后比较犹太教徒男性和女性的差异。研究者在调节变量的 3 个类别中进行了 3 次显著性检验。这些比较通常被称为"简单效应"或"简单主效应"。它们关注的是关键自变量在给定的调节变量类别中对结果变量的作用。以表 1.1 中的数据为例,天主教徒的性别效应是对男性与女性态度差异 (5.0－3.0＝2.0) 的显著性检验。

另一种看待简单效应的方法是把它们当做条件效应。一个自变量对结果变量的简单效应的条件是调节变量取某特定值(例如,调节变量取值为"犹太教")。

## 交互比较

简单主效应往往是研究者关心的问题，但它们与交互作用关系不大（尽管有些研究者认为它们有影响，稍后我们会讨论这一点）。对调节变量框架下的交互作用而言，关键自变量对结果变量的作用必须因调节变量取值的不同而变动。例如，天主教徒的性别差异必然与新教徒的性别差异不同，天主教徒的性别差异和犹太教徒的性别差异不同，新教徒的性别差异与犹太教徒的性别差异也不同。从表1.1的数据来看，天主教徒的性别差异是女性均值减去男性均值，即 $5.0-3.0=2.0$，新教徒的性别差异是 $6.0-3.0=3.0$，那么天主教徒的性别差异(2.0)就比新教徒的性别差异(3.0)小，这表明存在交互效应，即性别对堕胎态度的影响取决于宗教类别。这一效应可以通过计算两个均值差之间的差 $3.0-2.0=1.0$ 来反映。这一交互作用参数不为0。表明确实存在交互作用。当然，非零取值也可能只是反映抽样误差，所以有必要对估计参数进行正式的显著性检验。

上文展示了所谓的单自由度交互比较。之所以称之为“交互作用”，是因为它明确地比较了关键自变量在调节变量一个类别中与在另一个类别中，对因变量的相应作用。天主教徒的性别差异(2.0)正式地与新教徒的性别差异(3.0)形成了对比。这种比较的统计检验在分子上只有一个自由度，因此就被称为“单自由度的交互比较”。关于这种比较的更多讨论，请参考Jaccard(1996)。

单自由度的交互比较是交互分析的核心，它们代表了交互作用的关键检验。它们与整体交互检验不同，后者涉及一个以上的自由度。整体交互显著性检验是交互作用的总体检验，关注自变量和调节变量整体的而非其子群体的交互作用。在表1.1的析因设计中，整体检验关注的是性别和宗教之间总的交互作用，此例中自由度($df$)为2。相对而言，单自由度交互作用关注总样本中的 $2\times 2$ 子表。在实践中，研究者很少满足于只在总体层面作出陈述。通常，更多重要的问题都会采用单自由度的交互比较。

现在，我们明确说明了为何简单主效应不能解释统计交互作用的机制。交互比较正式地比较了一个自变量在调节变量的一个类别上与另一个类别上对

因变量影响的差异。相对而言,简单主效应并没有进行这种比较。简单主效应关注的只是调节变量在某一个类别上时,自变量是否在此特定类别中对因变量有影响。例如,对天主教徒而言,是否存在性别差异?简单效应的显著性检验并没有比较其他群体的影响(例如,它没有比较新教徒和犹太教徒的效应),它不能解决统计交互问题。

我们用另一种方法在一个相关的例子中来说明这一点。假设有两个变量,$X$ 和 $Y$,男性 $X$ 与 $Y$ 的相关系数是 0.24,女性的是 0.22。假设对男性而言,这一相关系数是显著的($p < 0.05$),但对女性是不显著的($p > 0.05$)。每组中简单效应的显著性检验是类似的。那我们能否从这些检验中得出结论说,$X$ 和 $Y$ 的相关性对男性来说比女性更强?当然不能。即使相关性在一组中统计显著,而在另一组不显著,我们也只能说,如果直接检验这两个相关性,那么它们存在差异。这种两个相关性差异的检验(在此例中并不统计显著)与调节变量框架下的交互作用的检验是类似的。

概言之,后面的章节对交互分析的讨论会考虑简单效应(也被称为"条件效应")、单自由度交互比较和整体交互检验。这几种检验都具有潜在的理论意义,在使用交互模型的多元回归方法中都会出现。我们会使用定性变量和简单析因设计来展示这些概念,也会在包含连续性解释变量的回归模型中考察相应的情况。

## 多元回归分析的回顾

假设读者熟悉多元回归的基础知识,本节的目的是介绍一些术语,并建构一个讨论问题的参考框架。多元回归的导论可以参考 Berry 和 Feldman (1985),或 Schroeder、Sjoquist 和 Stephan(1986)的著作。与传统方法一样,我们会使用希腊符号与字母来代表总体系数和总体数据,用阿拉伯数字来代表样本估计系数和样本数据。

### 线性模型

请考虑 3 个连续变量的例子,研究者关心两个自变量 $X_1$ 和 $X_2$ 对一个因变

量 $Y$ 的影响。分析样本数据主要采取的是最小二乘法方程：

$$Y = a + b_1X_1 + b_2X_2 + e$$

在这里，$a$ 代表最小二乘法估计的总体截距，$b_1$ 和 $b_2$ 是最小二乘法估计中的 $X_1$ 和 $X_2$ 的总体回归系数，$e$ 是残差。此方法要求总体数据结构满足如下条件，才能应用普通最小二乘法(OLS)并进行严格意义上的推论检验：(1)待检验线性模型适用于总体所有成员的真实模型；(2)总体残差独立分布，均值为 0；(3)解释变量本质上是固定的，有正的方差；(4)样本数据矩阵的阶等于列的数目，并少于观察值数量(例如，没有完全的多重共线性)；(5)残差在一组给定的 $X$ 取值下是正态分布的，且方差等于任何其他固定 $X$ 取值下残差的方差(即不存在异方差性)。当以上假定都得到满足时，OLS 估计就被称为"最佳线性无偏估计"，因为它是线性的、无偏的，且具有一组无偏估计值中最小的方差。放松某些假定只会给推论结果和参数估计带来较小的影响，而违反某些假设则会带来问题。例如，尽管许多研究实践依赖于解释变量是随机的、非固定的，但 OLS 依然是有效的，只需研究者假定随机解释变量是以观察到的 $X$ 为条件的一个真实样本。第 4 章会深入讨论违反假定的情况。

样本的复相关系数 $R$ 是样本数据中模型整体拟合程度的指标。回归系数则通常被理解为控制方程中其他 $X$ 变量时，某 $X$ 变量对 $Y$ 的影响。具体而言，某个特定的 $b$ 代表当所有其他 $X$ 都保持不变时，$X$ 变化一个单位，$Y$ 可能变化的单位。如果变量被标准化，那么截距一直为 0，而 $b$ 就代表了标准化回归系数；对系数的理解是一样的，只不过单位变成了标准分数。[1]

因此，如果解释变量 $X$ 的标准化回归系数 $b$ 为 1.5，就表示 $X$ 每增加一个标准分，$Y$ 将会变动 1.5 个标准分。有些社会科学家在多元回归中更倾向于使用标准化回归系数，而不是非标准化回归系数，因为标准化之后，所有变量都被认为具有共同的度量，更容易实质性地比较不同自变量系数的大小。第 4 章会讨论这一问题的细节。

请考虑下面这个例子。某个社会学家对此问题感兴趣：通过 6 类关系的满意或不满意来预测人们对自己婚姻关系的满意程度。共有 339 个人给出其对婚

姻的满意或不满意分数，为 11 分制，从 −5 到 +5，表示从非常不满意到非常满意（分数越高，满意度越高）。另外，他们还给出了对其生活中 6 个方面关系的满意程度（同样是 11 分制）：交往数量、情感表达方式、情感支持数量、共同爱好层次、共处时间以及争议解决方式。调查者进行多元回归分析，用 6 个因子对婚姻满意程度进行回归。表 1.2 展示了 SPSS 软件的电脑输出结果。

复相关系数平方是 0.663，这表示总体婚姻满意度的方差能被六因素线性组合所解释的比例。虚无假设指的是，总体的复相关系数为 0。通过 $F$ 检验来检验这一假设，结果在表的上方。$F$ 值是统计显著的 [$F(6,332) = 108.70$]，这就拒绝了虚无假设。标准化和非标准化的回归系数在表 1.2 的下方。如上所述，非标准化回归系数反映了其他 $X$ 变量不变时，$X$ 变动一个单位导致总体婚姻满意度的变动。例如，情感支持满意度每增加一个单位，总体婚姻关系满意度会增加 0.307 个单位。相比之下，共处时间每增加一个单位，只能使总体满意度增加 0.005 个单位，当然，这都是在控制其他变量不变的情况下。标准化回归系数的理解是类似的，不过不是在原始分数，而是在标准分数的意义上。

每个非标准化回归系数都有一个标准误（在"SE/标准误"那一栏）。这些统计数字代表估计总体回归系数时抽样误差的大小。具体而言，估计的标准误表明针对随机样本大小为 $N$ 的所有抽样，样本估计值与总体真实参数值之间的平均偏离水平。当其他条件不变时，标准误越大，抽样误差就越大，我们对样本估计也就越没有信心。

标有"$t$"的那栏代表虚无假设的 $t$ 检验，即某一特定的回归系数等于 0。$t$ 值等于回归系数除以估计值的标准误。标有"Sig."的那栏表示的是 $t$ 检验的 $p$ 值。在上例中，除（婚姻满意度与）共处时间满意度的相关性之外，其他的回归系数都是统计显著的（$p < 0.05$）。

最后，每一解释变量与标杆（指因变量）的相关性还可通过解释变量与因变量的零阶相关性（见"Zero-order"栏），以及解释变量和因变量的偏相关性来展示（控制住其他解释变量的影响，见"Part"栏）。前一统计值（零阶相关系数）的平方反映的是当所有其他的满意因素都允许自由变动时，某解释变量能够解释的总体婚姻满意度的比例。而后一个统计值（偏相关系数）的平方反映的是当其他变量

都被控制时，这一变量单独能够解释的因变量方差的比例。例如，情感支持满意度能够解释总体婚姻满意度方差的比例是 $(100)\times(0.72)\times(0.72)=51.5\%$，而它能够单独解释的因变量方差的比例是 $(100)\times(0.24)\times(0.24)=5.8\%$（其他的部分由另外 5 个解释变量所解释）。[2]

## 层级回归

研究者通常进行多元层级回归。在此情况下，研究者对加入一个或更多解释变量能否显著增加对因变量的解释力感兴趣。通常额外解释的方差是这样计算的：用原方程的复相关系数平方减去扩展方程的复相关系数平方。两个复相关系数平方的差异就是额外解释变量所能解释的因变量的方差。

例如，如果差异等于 0.10，那么通过引入新的解释变量可额外解释因变量方差的 10%。可通过下面这个方程检验复相关系数平方在总体中为 0 的虚无假设：

$$F=\frac{(R_2^2-R_1^2)/(k_2-k_1)}{(1-R_2^2)/(N-k_2-1)} \qquad [1.1]$$

在这里，$R_2$ 代表的是扩展方程的复相关系数，而 $R_1$ 则是原方程的复相关系数；$k_2$ 是扩展方程的解释变量数目，而 $k_1$ 是原方程的解释变量数目；$N$ 则是总样本量。$F$ 值服从 $k_2-k_1$ 和 $N-k_2-1$ 个自由度的 $F$ 分布。

如上例，层级检验常被应用在向基本模型中增加一个解释变量的这种情况。然而，有时它会扮演从方程中删减变量的角色。假设回归方程有 6 个解释变量，我们希望判断去掉其中两个变量对复相关系数的影响。两个方程的复相关系数平方的变化反映了去掉两个变量时，方程能解释的因变量方差比例的变化。我们利用方程 1.1 来考察去掉两个变量后，模型所能解释方差的变化的统计显著性。这里，$R_2$ 表示解释变量较多的模型的复相关系数，而 $R_1$ 代表解释变量较少的模型的复相关系数。

## 分类解释变量和虚拟变量

回归分析通常包含分类变量作为解释变量，例如性别、民族、宗教信仰。这

**表 1.2　SPSS 多元回归输出结果**

| 模　型 | $R$ | $R^2$ | 调节的 $R^2$ | 测得 SE |
|---|---|---|---|---|
| 1 | 0.814 | 0.663 | 0.457 | 0.74094 |

| | 平方和 | $df$ | 平方均值 | $F$ | Sig. |
|---|---|---|---|---|---|
| 回归系数 | 357.897 | 6 | 59.650 | 108.698 | 0.000 |
| 残　差 | 182.191 | 332 | 0.549 | | |
| 总　计 | 540.088 | 338 | | | |

| 模　型 | 非标准化回归系数 | | 标准化回归系数 | $t$ | Sig. | B 的 95% 置信区间 | | 相关系数 | |
|---|---|---|---|---|---|---|---|---|---|
| | B | SE | β | | | 低端 | 高端 | Zero-order | Part |
| （常数项） | 1.010 | 0.127 | | 7.980 | 0.000 | 0.761 | 1.259 | | |
| 争议解决 | 0.121 | 0.025 | 0.191 | 4.880 | 0.000 | 0.072 | 0.170 | 0.578 | 0.155 |
| 共处时间 | 0.005 | 0.023 | 0.008 | 0.220 | 0.823 | −0.044 | 0.054 | 0.353 | 0.007 |
| 情感表达 | 0.151 | 0.039 | 0.173 | 3.910 | 0.000 | 0.075 | 0.227 | 0.629 | 0.210 |
| 共同爱好 | 0.153 | 0.034 | 0.183 | 4.460 | 0.001 | 0.086 | 0.220 | 0.550 | 0.142 |
| 交往数量 | 0.125 | 0.037 | 0.147 | 3.330 | 0.000 | 0.058 | 0.192 | 0.625 | 0.106 |
| 情感支持 | 0.307 | 0.041 | 0.350 | 7.560 | 0.000 | 0.227 | 0.387 | 0.724 | 0.241 |

些变量在方程中用虚拟变量来表示。分析者为确定属于哪一群组而生成一个虚拟变量。以性别变量为例，生成一个虚拟变量，然后所有男性赋值为 1，所有女性赋值为 0。这种赋值方法称为“虚拟编码”，即所有属于某一群组的个案赋值为 1，所有不属于这一群组的个案赋值为 0。当一个定性变量有两个以上的类别时，必须设定一个以上的虚拟变量来表示不同的群组。一般而言，我们需要 $m-1$ 个虚拟变量，这里的 $m$ 指的是一个变量的类别。

假设有一个解释变量——党派，有 3 个取值——民主党、共和党或独立党。在此例中，需要 $3-1=2$ 个虚拟变量来代表党派。第一个虚拟变量是 $D_{\mathrm{D}}$，所有的民主党党员都赋值为 1，其他人都赋值为 0。第二个虚拟变量是 $D_{\mathrm{R}}$，共和党党员赋值为 1，其他人都赋值为 0。尽管可以产生第三个虚拟变量来代表独立党，并且独立党党员赋值为 1，其他人赋值为 0，但这一变量与另外两个虚拟变量放在一起是冗余的。一旦知道某人是否为民主党党员，且知道某人是否为共和党党员(通过第一个和第二个虚拟变量的值)，根据定义就自然知道此人是否为独立党党员。这一点背后的逻辑对性别变量来说更明显。生成一个虚拟变量来区分这两个群体，男性赋值为 1，女性赋值为 0。如果产生第二个虚拟变量，女性赋值为 0，男性赋值为 1，那么这个虚拟变量就与第一个虚拟变量完全负相关，因此是冗余的。通过虚拟编码，那些在任一虚拟变量上都未被赋值为 1 的个案组成了参照组。在上例中，性别的参照组是女性；而在党派例子中，参照组是独立党。从统计角度来讲，可任意选择一组作为参照组。

假设用对堕胎态度的一个测度对宗教类别这一分类变量进行回归。态度变量的取值是 0 到 100，分数越高表明越支持。宗教变量有 3 组，天主教、新教和犹太教，它由两个虚拟变量来表示，一个虚拟变量是所有天主教徒赋值为 1，其他教派为 0($D_{\mathrm{C}}$)；另一个是所有的新教徒赋值为 1，而其他人赋值为0($D_{\mathrm{P}}$)。犹太教徒是参照组。假设回归分析的结果是复相关系数平方 0.30，这表明宗教能够解释堕胎态度 30％的方差。针对复相关系数平方的显著性检验——$F$ 检验，检验的虚无假设是宗教对堕胎态度没有影响，或 3 个宗教群体的堕胎态度均值无差异。两个虚拟变量的非标准化回归系数很容易理解，每个系数都代表了一个平均的差异。具体而言，系数代表这个取值为 1 的组的平均得分与参照组

平均得分之差。假设 $D_C$ 的回归系数是－1.0，这表明天主教徒态度分数均值减去犹太教徒的均值是－1.0，或者说，平均而言，天主教徒的堕胎态度分数比犹太教徒的堕胎态度分数低 1.0 个单位。$D_P$ 的回归系数是－0.5，这表明，平均而言，新教徒的堕胎态度分数比犹太教徒的堕胎态度分数低 0.5 个单位。这些回归系数的显著性检验，检验的是这些平均差异的显著程度。

如果重复上面的分析，但加入一个连续变量到回归方程中，那么虚拟变量的回归系数依然表示平均差异，不过这是模型中其他变量保持不变的情况下的平均差异(例如，上述连续变量)。例如，在方程中加入社会阶级这一测度，则 $D_C$ 的回归系数变成了－0.80，这就是在控制社会阶级变量时，天主教徒和犹太教徒堕胎态度的平均差异。

还有许多给虚拟变量赋值的方法。如上所述，依靠 1 和 0 的赋值方法被称为“虚拟编码”。Hardy 讨论了不同编码方案的逻辑(Hardy，1993)。虚拟编码很有用，因为它很好地与交互作用的调节变量概念匹配，在下面的章节中会分析这一点。

## 多元回归中的预测值

假设一组数据分析的是性别和用保守主义—自由主义来测量的意识形态对一个环境组织提倡的清理河道运动的捐款数目(单位为美元)的影响。意识形态测度的取值范围是－3 到＋3，0 代表中立点，负值代表程度更高的保守主义，而正值代表程度更高的自由主义；性别用虚拟变量来代表，男性取值为 1，女性取值为 0；结果变量的单位是美元。假设分析得到下面这个回归方程：

$$Y' = 10.00 + (-5.50)Gender + 1.00Ideology \qquad [1.2]$$

这里的 $Y'$ 是预测的个体捐款金额。给定一个解释变量的取值，把这个值代入方程中，就可计算 $Y$ 的预测值。例如，对一个意识形态分数为＋2 的男性来说，其捐款金额预测值就是：

$$Y' = 10.00 + (-5.50) \times 1 + 1.00 \times (2) = 6.50$$

对那些意识形态分数为－2 的男性，其捐款金额的预测值又是多少呢？通过代

换，我们得到：

$$Y' = 10.00 + (-5.50) \times 1 + 1.00 \times (-2) = 2.50$$

在后面的分析中，我们将会用到这些预测值。

## 解释变量的变换

通过在回归之前对解释变量进行代数处理，可使回归系数直接反映那些有理论意义的参数。后面的章节会揭示这么做的好处，但在此，我们会阐释基本的逻辑。上例中，假定在回归分析之前，把意识形态变量减去一个常数 1。原先尺度的取值范围是－3 到＋3，新尺度的取值范围变成了－4 到＋2，每个分数都往下平移一个单位。运用这种变换分数的回归分析结果如下：

$$Y' = 11.00 + (-5.50)\mathit{Gender} + 1.00\mathit{Ideology}_{\mathrm{t}}$$

请注意，只有一个参数受到这种转换的影响，即截距，而其他参数与原先的分析结果并无二致。截距表示性别为 0 且变换后意识形态分数也为 0 时的预测值。但在这里，变换后的意识形态分数 0，代表原先的意识形态分数为＋1。第二个分析的截距应当等于原先的分析中意识形态分数为 1 的女性的预测值。事实正是如此。在原方程中：

$$Y' = 10.00 + (-5.50) \times 0 + 1.00 \times (1) = 11.00$$

这与第二次分析中的截距是相同的。为何要进行这种变换呢？几乎所有的电脑程序都不仅报告回归方程中的参数估计，还给出估计值的标准误和给定估计的置信区间。进行如上变换代表了一种简单的、计算任意解释变量组合的 $Y$ 预测值置信区间的方法。简单变换每一个解释变量，通过增加或减去一个常数项，使变换后的 0 值代表原先尺度中研究者感兴趣的解释变量的取值。利用变换后的解释变量进行回归，方程的截距项就提供了自变量特定组合下预测的 $Y$ 均值，且截距项的置信区间就是与此自变量组合相连的 $Y$ 预测值的置信区间。

如果进行并非这种性质的变换，那么很多时候截距项就不易理解，因为它

反映了 $Y$ 预测值的均值，但解释变量取值为 0 的情况并不存在，或者说超出了研究变量的取值范围。后面的章节还会用到这一变换策略及其变形，这是用来分离一系列显著性检验和置信区间的关键策略。

## 本书概览

本书余下的部分会讨论一系列与交互分析相关的议题。我们的关注点在于连续变量的分析以及连续变量和定性变量混在一起时的分析，因为精妙的交互分析同样出现在所有解释变量都是定性变量的情况下（例如，Cohen & Cohen，1983）。第 2 章关注双向交互作用，并以传统的乘积项分析作为双向线性交互作用的分析工具。第 3 章考察三向交互作用。第 4 章关注各种已经整合到交互分析文献中的话题，并协助应用型研究者在运用交互模型时考虑一些重要的问题。

# 第 2 章 | 双向交互作用

本章首先考察包含交互作用的多元回归模型设定的一般性问题。其次考察两个解释变量都是连续变量时的双向交互作用。最后，我们讨论一个解释变量是定性的，而另一个是连续变量的情况。

## 包含乘积项的回归模型

举一个有 3 个连续变量的例子：研究者对两个自变量（$X$ 和 $Z$）对一个因变量（$Y$）的影响感兴趣。具体来说，假设研究者试图理解为何发生性行为时，有些青少年不采取避孕措施，而有些青少年却倾向于采取避孕措施。研究者获取了一个包含 125 名有性生活的女青年样本，并获取每个女青年采取避孕措施意愿的测度。这一测度包含一个评分尺度，一端是“绝对不愿采取避孕措施”，另一端是“绝对希望采取避孕措施”。分数的取值是 0 到 30，分数越高代表采取避孕措施的意愿越强。研究者假设有两组因素会影响这种意愿：第一组是个体对采取避孕措施的感受或者态度，即个体喜欢还是不喜欢避孕措施；第二组是感受到采取避孕措施的同侪压力。所有这些因素，态度和感受到的同侪压力，都用 1 到 5 的 5 分制尺度来衡量。对态度测度而言，分数越高，个体越倾向于采取避孕措施；对同侪压力测度而言，分数越低，代表越少的采取避孕措施的同侪压力，而分数越高则代表这种压力水平在上升。

如第 1 章指出的，检验加叠性（或“主效应”）模型中 $X$ 和 $Z$ 解释 $Y$ 的显著性，通常采取最小二乘法方程，总体模型如下：

$$Y = \alpha + \beta_1 X + \beta_2 Z + \varepsilon \qquad [2.1]$$

应用方程 2.1 涉及避孕意愿对态度和同侪压力进行回归。假设研究者对是否存在交互效应感兴趣。具体而言，研究者假设态度和意愿之间的关系受到同侪压力的调节作用：当同侪压力最小时，个体态度就会对意愿产生一个更强的影响；当同侪压力很大时，态度的影响力就小了。在此例中，关键自变量是使用避孕措施的态度($X$)，调节变量是同侪压力($Z$)。

最常见的处理回归分析中交互作用的方法是使用乘积项。我们可以用简单的代数来展现乘积项的基本逻辑。如果态度对意愿的影响通过方程 2.1 中的 $\beta_1$ 来反映，且 $Z$ 被认为对态度对意愿的影响有调节作用，那么当 $Z$ 变化时，$\beta_1$ 的取值也会相应地变化。可用下面这个线性方程说明 $\beta_1$ 为何可能是 $Z$ 的函数：

$$\beta_1 = \alpha' + \beta_3 Z \qquad [2.2]$$

根据这个方程，$Z$ 每变化一个单位，$\beta_1$ 会被预计有 $\beta_3$ 个单位的变化。现在，把右边部分代入方程 2.1 的 $\beta_1$ 项，得到的结果是：

$$Y = \alpha + (\alpha' + \beta_3 Z)X + \beta_2 Z + \varepsilon$$

将 $X$ 乘入，得到的结果是：

$$Y = \alpha + \alpha' X + \beta_3 ZX + \beta_2 Z + \varepsilon$$

那么，给系数一些新标签并调整顺序，我们就可以得到包含乘积项的交互模型：

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 ZX + \varepsilon \qquad [2.3]$$

这一阐述过于简单，因为其他类型的交互模型会产生同样的方程，而其他类型的交互模型可以从不同的交互模型中演化而来。在此，只是给读者提供乘积项背后的逻辑来反映关键自变量对结果变量的作用被认为是调节变量线性函数的情况。

## 两个连续性解释变量

### 传统的回归策略

我们继续讨论那个用态度和同侪压力来预测(避孕)意愿的交互作用分析

的例子。数据显示在表 2.1 中。这些数据是虚构的,用于反映非真实的系统性的属性,以揭示下面讨论中涉及的问题。出于教学的考虑,数据根据析因表,即每一个 $X$ 和 $Z$ 的组合给出 $Y$ 相应的均值分数,这体现在表 2.2 中。这个表格本质上是 5×5 的析因设计。

多元回归中最流行的分析交互作用的策略是由 Cohen 等人(1983)提出的。这涉及一个乘积项,$XZ$,它被认为包含了交互作用,然后计算两个 $R^2$ 的值,一个模型"只含主效应"(方程 2.1),另一个模型在主效应模型中加入乘积项,从而这组样本数据产生了一个有 3 个解释变量的方程:

$$Y = a + b_1 X + b_2 Z + b_3 XZ + e \qquad [2.4]$$

如果存在交互作用,那么两个 $R^2$ 的差异应当是统计显著的(除非有类型 1 错误)。[3] 可用方程 1.1 中的层级 $F$ 检验对这种差异进行正式的显著性检验。

使用不包括乘积项的标准回归程序分析表 2.1 中的数据,得到两解释变量模型的复相关系数 $R$ 是 0.90139。回归方程是:

$$Y = 8.0 + 3.0X + (-2.0)Z + e \qquad [2.5]$$

三项交互模型的复相关系数 $R$ 是 0.96825,回归方程是:

$$Y = -1.0 + 6.0X + 1.0Z + (-1.0)XZ + e \qquad [2.6]$$

应用方程 1.1 得到下面这个结果:

$$F = \frac{(0.96825^2 - 0.90139^2)/(3-2)}{(1-0.96825^2)/(125-3-1)} = 242.26$$

**表 2.1 交互作用的虚构数据**

| *ID* | *Y* | *X* | *Z* | *ID* | *Y* | *X* | *Z* | *ID* | *Y* | *X* | *Z* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 1 | 1 | 7 | 4 | 1 | 2 | 13 | 5 | 1 | 3 |
| 2 | 4 | 1 | 1 | 8 | 5 | 1 | 2 | 14 | 6 | 1 | 3 |
| 3 | 5 | 1 | 1 | 9 | 6 | 1 | 2 | 15 | 7 | 1 | 3 |
| 4 | 6 | 1 | 1 | 10 | 7 | 1 | 2 | 16 | 3 | 1 | 4 |
| 5 | 7 | 1 | 1 | 11 | 3 | 1 | 3 | 17 | 4 | 1 | 4 |
| 6 | 3 | 1 | 2 | 12 | 4 | 1 | 3 | 18 | 5 | 1 | 4 |

续表

| *ID* | *Y* | *X* | *Z* | *ID* | *Y* | *X* | *Z* | *ID* | *Y* | *X* | *Z* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 6 | 1 | 4 | 55 | 17 | 3 | 1 | 91 | 9 | 4 | 4 |
| 20 | 7 | 1 | 4 | 56 | 11 | 3 | 2 | 92 | 10 | 4 | 4 |
| 21 | 3 | 1 | 5 | 57 | 12 | 3 | 2 | 93 | 11 | 4 | 4 |
| 22 | 4 | 1 | 5 | 58 | 13 | 3 | 2 | 94 | 12 | 4 | 4 |
| 23 | 5 | 1 | 5 | 59 | 14 | 3 | 2 | 95 | 13 | 4 | 4 |
| 24 | 6 | 1 | 5 | 60 | 15 | 3 | 2 | 96 | 6 | 4 | 5 |
| 25 | 7 | 1 | 5 | 61 | 9 | 3 | 3 | 97 | 7 | 4 | 5 |
| 26 | 8 | 2 | 1 | 62 | 10 | 3 | 3 | 98 | 8 | 4 | 5 |
| 27 | 9 | 2 | 1 | 63 | 11 | 3 | 3 | 99 | 9 | 4 | 5 |
| 28 | 10 | 2 | 1 | 64 | 12 | 3 | 3 | 100 | 10 | 4 | 5 |
| 29 | 11 | 2 | 1 | 65 | 13 | 3 | 3 | 101 | 23 | 5 | 1 |
| 30 | 12 | 2 | 1 | 66 | 7 | 3 | 4 | 102 | 24 | 5 | 1 |
| 31 | 7 | 2 | 2 | 67 | 8 | 3 | 4 | 103 | 25 | 5 | 1 |
| 32 | 8 | 2 | 2 | 68 | 9 | 3 | 4 | 104 | 26 | 5 | 1 |
| 33 | 9 | 2 | 2 | 69 | 10 | 3 | 4 | 105 | 27 | 5 | 1 |
| 34 | 10 | 2 | 2 | 70 | 11 | 3 | 4 | 106 | 19 | 5 | 2 |
| 35 | 11 | 2 | 2 | 71 | 5 | 3 | 5 | 107 | 20 | 5 | 2 |
| 36 | 6 | 2 | 3 | 72 | 6 | 3 | 5 | 108 | 21 | 5 | 2 |
| 37 | 7 | 2 | 3 | 73 | 7 | 3 | 5 | 109 | 22 | 5 | 2 |
| 38 | 8 | 2 | 3 | 74 | 8 | 3 | 5 | 110 | 23 | 5 | 2 |
| 39 | 9 | 2 | 3 | 75 | 9 | 3 | 5 | 111 | 15 | 5 | 3 |
| 40 | 10 | 2 | 3 | 76 | 18 | 4 | 1 | 112 | 16 | 5 | 3 |
| 41 | 5 | 2 | 4 | 77 | 19 | 4 | 1 | 113 | 17 | 5 | 3 |
| 42 | 6 | 2 | 4 | 78 | 20 | 4 | 1 | 114 | 18 | 5 | 3 |
| 43 | 7 | 2 | 4 | 79 | 21 | 4 | 1 | 115 | 19 | 5 | 3 |
| 44 | 8 | 2 | 4 | 80 | 22 | 4 | 1 | 116 | 11 | 5 | 4 |
| 45 | 9 | 2 | 4 | 81 | 15 | 4 | 2 | 117 | 12 | 5 | 4 |
| 46 | 4 | 2 | 5 | 82 | 16 | 4 | 2 | 118 | 13 | 5 | 4 |
| 47 | 5 | 2 | 5 | 83 | 17 | 4 | 2 | 119 | 14 | 5 | 4 |
| 48 | 6 | 2 | 5 | 84 | 18 | 4 | 2 | 120 | 15 | 5 | 4 |
| 49 | 7 | 2 | 5 | 85 | 19 | 4 | 2 | 121 | 7 | 5 | 5 |
| 50 | 8 | 2 | 5 | 86 | 12 | 4 | 3 | 122 | 8 | 5 | 5 |
| 51 | 13 | 3 | 1 | 87 | 13 | 4 | 3 | 123 | 9 | 5 | 5 |
| 52 | 14 | 3 | 1 | 88 | 14 | 4 | 3 | 124 | 10 | 5 | 5 |
| 53 | 15 | 3 | 1 | 89 | 15 | 4 | 3 | 125 | 11 | 5 | 5 |
| 54 | 16 | 3 | 1 | 90 | 16 | 4 | 3 | | | | |

表 2.2 作为 $X$ 和 $Z$ 函数的单元格均值

| 态度($X$) | 同侪压力($Z$) | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 5 | 25 | 21 | 17 | 13 | 9 |
| 4 | 20 | 17 | 14 | 11 | 8 |
| 3 | 15 | 13 | 11 | 9 | 7 |
| 2 | 10 | 9 | 8 | 7 | 6 |
| 1 | 5 | 5 | 5 | 5 | 5 |

对 1 和 121 的自由度($F$ 检验)而言,$F$ 值是统计显著的,表明存在统计交互作用。层级 $F$ 检验产生的 $p$ 值与方程 2.4 中 $b_3$ 的显著性检验,即 $t$ 检验的 $p$ 值应当相同。在此样本中,$b_3$ 的 $t$ 值是 15.56,而这个值的平方正是大家所看到的 $F$ 值,即 $15.56 \times 15.56 = 242.26$。因此,与其执行层级检验来决定交互作用的统计显著性,不如简单地检验 $b_3$ 的统计显著性。

下面挑选一些有关这一分析的重要问题进行讨论。我们用避孕措施的例子来讨论这些问题,随后再用新的例子来展示研究实践中的应用。

## 交互的形式

如上所述,简单的乘积项允许研究者检验是否存在调节关系。原则上来说,有一系列调节关系可用于刻画两个连续变量之间的交互作用。例如,$X$ 对 $Y$ 作用的斜率是 $Z$ 的线性函数。这被称为“双线性交互作用”,表 2.2 中的关系正是如此。当同侪压力低至 1 时,意愿对态度的斜率是较高的,即 5.0。态度每变化一个单位,意愿会改变 5.0 个单位。当同侪压力增加时,这一斜率就降低了。例如,当同侪压力为 3 时,意愿对态度的斜率变化为 3.0,即态度每变化一个单位,意愿对态度的斜率只变化 3 个单位。每当同侪压力增加一个单位,这个斜率就降低一个单位。斜率变化和调节变量变化之间这种顺序的、单调的线性关系,正是传统乘积项方法下统计交互作用的精华。

其他类型的函数关系也是可能的。例如,当某人从低同侪压力向中同侪压力变动时,意愿—态度斜率的变化更大。然而,当某人从中同侪压力向高同侪

压力转变的时候，斜率变化的程度会稍稍平缓，直到它们达到一个最小的变化点。另外，意愿和态度之间的关系可能是非线性的。自变量和因变量关系的曲线的形状（而非斜率）可能随着同侪压力的变动而变动。涉及连续性解释变量的可能函数关系的数量是无限的。在此需要强调，我们只使用了一种函数形式，即双线性交互作用，如方程 2.4 中的简单乘积项所示。若使用传统的乘积项却发现交互作用统计不显著，那么就说明可能存在另一种函数形式（的交互作用），而并非一定不存在调节关系。第 4 章会讨论识别交互作用其他形式的策略以及检验它们的策略。在理想情况下，理论应当预测一个特定的函数形式，然后，数据分析应当建构适当的模型来检验函数的形式。本节余下的部分将分析局限于双线性交互作用。

### 理解乘积项的回归系数

方程 2.4 中所有的回归系数都需要解释。我们首先考虑 $b_3$，乘积项的回归系数。$b_3$ 是单自由度的交互比较，传达了交互作用的本质信息。它表明 $Z$ 增加一个单位，$X$ 对 $Y$ 作用的斜率所变化的单位数量。在我们的样本中，$b_3$ 为 $-1.0$，这表示同侪压力每增加一个单位，态度对意愿作用的斜率就降低一个单位。这可以从表 2.2 中看出来。

表 2.2 的数据是虚构的，在给定 $Z$ 值的情况下，$X$ 对 $Y$ 作用的斜率可通过对数据的观察而发现。然而，研究实践中很少有这种情况。当使用 3 个回归系数可计算任意给定的 $Z$ 值时，$X$ 对 $Y$ 的作用只需一些代数计算。

当要分析 $Y$ 和 $X$ 的关系时，必须先设定 $Z$（同侪压力）的取值。让我们从 $Z$ 的最小可能取值 1 开始。利用三项式回归方程，把 1 代入 $Z$ 的位置。得到下列方程：

$$Y = -1.0 + 6.0X + (1.0)(1) + (-1.0)(X)(1) + e$$

重新安排方程右边的项目，合并同类项，$X$ 仍然在右边：

$$Y = -1.0 + (1.0)(1) + 6.0X + (-1.0)(X)(1) + e$$

接着把 $X$ 的相关项合并：

$$Y^{[4]} = -1.0 + (1.0)(1) + [6.0 + (-1.0)(1)](X) + e$$

得到：

$$Y = 0.0 + 5.0X + e$$

这一结果是指当 $Z$ 取 1 时，$Y$ 和 $X$ 之间关系的线性方程。接下来进行同样的计算，把 $Z$ 的最大值 5 代入，得到的方程是：

$$Y = -1.0 + 6.0X + (1.0)(5) + (-1.0)(X)(5) + e$$

进行上面同样的计算，得到下面的结果：

$$Y = 4.0 + 1.0X + e$$

这些计算揭示了 $Y$ 和 $X$ 之间的关系如何在 $Z$ 的取值范围内变动：在 $Z$ 值的低端（比如，1），$X$ 一个单位的变动能使 $Y$ 的预测值变动 5 个单位；而在 $Z$ 值的高端（比如，5），$X$ 一个单位的变动使 $Y$ 的预测值变动 1 个单位。可用一个方程来计算在 $Z$ 的每个取值上，$X$ 对 $Y$ 预测值的作用，它就是：

$$b_1 \text{ 在 } Z = b_1 + b_3 Z \qquad [2.7]$$

以我们的例子来说，当 $Z$ 的取值为 4 时，$X$ 对 $Y$ 的作用斜率如下：

$$b_1 \text{ 在 } 4 = 6.0 + (-1.0)(4) = 2.0$$

若计算在 $Z$ 的每一个取值上，$X$ 取值对 $Y$ 作用的斜率，那么得到：

$$b_1 \text{ 在 } Z \text{ 取值为 } 1 = 5.0$$

$$b_1 \text{ 在 } Z \text{ 取值为 } 2 = 4.0$$

$$b_1 \text{ 在 } Z \text{ 取值为 } 3 = 3.0$$

$$b_1 \text{ 在 } Z \text{ 取值为 } 4 = 2.0$$

$$b_1 \text{ 在 } Z \text{ 取值为 } 5 = 1.0$$

请注意，$Z$ 每增加一个单位，斜率取值就降低 1.0 个单位（$-1.0$），这正是 $b_3$ 的值。另外，乘积项系数取值揭示了调节变量变动一个单位时，斜率随之变动的单位数量。

## 理解组成项的回归系数

有些社会科学家强调，在一个方程中加入乘积项会使得原组成项（$X$ 和 $Z$）的回归系数较难理解。研究者注意到，相比那些只包含主效应方程的相应系数，加入乘积项往往会使这些系数发生剧烈变动。事实上，系数甚至会改变方向。这种“矛盾”使在包含乘积项模型中理解回归系数变得困难。

事实上，三项式方程中的回归系数需要我们去解释。上面提到的这些差异是因为两个方程估计的是不同的概念。在两项只含主效应的模型中，回归系数估计一个解释变量对因变量的影响，并控制其他解释变量的每一个层次。在两项式模型中，$b_1$ 反映的是在每一个 $Z$ 值上，$X$ 变动一个单位，$Y$ 随之变动的单位；而 $b_2$ 反映的是在每个 $X$ 值上，$Z$ 变动一个单位，$Y$ 变动的单位数。对包含乘积项的模型而言，$X$ 和 $Z$ 的回归系数反映了基于特定 $Z$ 和 $X$ 值条件下的关系：$b_1$ 反映的是当 $Z$ 等于 0 时，$X$ 对 $Y$ 的影响；$b_2$ 反映的是当 $X$ 等于 0 时，$Z$ 对 $Y$ 的影响。在两个模型中，$b_1$ 和 $b_2$ 的系数的差异来自“主效应”模型，这两个系数反映了其他解释变量取平均水平时，$X$ 和 $Z$ 的“通常”关系；在乘积项模型中，它们反映了关注包含乘积项其他解释变量特定取值时的条件关系。[5]

换句话说，在乘积项模型中，$X$ 和 $Z$ 的回归系数并非如传统认为的那样代表主效应。相反，这些系数代表的是简单效应。$X$ 的系数估计的是 $Z$ 为特定值（即 $Z=0$）时，$X$ 对 $Y$ 的作用，而 $Z$ 的系数估计的是 $X$ 为特定值（即 $X=0$）时，$Z$ 对 $Y$ 的影响。有些研究者错误地理解了这些系数，把它们当成主效应，实际上它们不是。我们必须牢记这些系数的本质。

仍以避孕研究为例，$X$ 的系数代表的是当同侪压力 $Z$ 为 0 时，态度 $X$ 对避孕意愿 $Y$ 的影响。系数为 6.0，因此当同侪压力对应的 $Z$ 值为 0 时，态度变化一个单位，预计会使意愿变化 6 个单位。

然而，这种理解有一个问题。系数表明了当 $Z=0$ 时，$X$ 对 $Y$ 的影响大小，但同侪压力为 0 的取值是不存在的。$Z$ 值的范围是 1 到 5，所以当 $Z=0$ 时，态度对意愿的影响其实没有实质意义。解决该问题的一个方法是变化 $Z$，使 0 值有意义。例如，$Z$ 的均值是 3.0，假设我们把每个样本的 $Z$ 的分数都减去 3.0，这

样，与原取值范围 1 到 5 不同，变化后的 $Z$ 值($Z_t$)的取值范围将是 $-2$ 到 $+2$。相比 $Z$ 的均值为 3.0，$Z_t$ 的均值将是 0。一个变量减去自身均值以形成新尺度的方法被称为“对中”。

假设我们对中 $Z_t$，重新计算乘积项 $XZ_t$，然后估计回归方程：

$$Y = a + b_1X + b_2Z_t + b_3XZ_t + e \qquad [2.8]$$

原来使用 $Z$ 作为解释变量的方程是：

$$Y = 1.0 + 6.0X + (1.0)Z + (-1.0)XZ + e$$

现在，$Z_t$ 取代了 $Z$ 的位置，方程变为：

$$Y = 2.0 + 3.0X + 1.0Z_t + (-1.0)XZ_t + e$$

这两个分析有几个特点。首先，利用 $X$、$Z$ 和 $XZ$ 来预测 $Y$ 的复相关系数平方及其显著性与利用 $X$、$Z_t$ 和 $XZ_t$ 是一致的。复相关系数(及其显著性检验)不随变换而变动。其次，$b_3$ 的值和它的显著性检验同样不随变换而改变。两个交互作用分析体现了完全相同的特征。尽管如此，请注意，$X$ 系数从原分析中的 6.0 变成了 3.0。在新的分析中，$X$ 系数是当 $Z_t = 0$ 时，$X$ 对 $Y$ 影响作用的大小。在此分析中，$Z_t$ 为 0 对应原来的 $Z$ 为 3。所以这一系数实际上反映的是 $Z = 3$ 时，$X$ 对 $Y$ 的影响大小。换句话说，当同侪压力中等时，若 $Z$ 值为 3，那么 $X$ 对 $Y$ 的影响大小反映为 $Y$ 对 $X$ 回归的斜率 3.0。通过变换 $Z$ 值，即通过使 0 值有意义，与 $X$ 相连的系数也变得有意义了。在新的分析中，$b_1$ 就是当 $Z$ 等于其样本均值时，$X$ 对 $Y$ 的预测作用大小。

现在给定交互模型中的一个 $Z$ 值，我们用另外两种方法来计算 $X$ 对 $Y$ 的作用。第一个策略是利用方程 2.7，我们可在此重复一下：

$$b_1 \text{ 在 } Z = b_1 + b_3Z$$

例如，当 $Z = 2$ 时，$b_1$ 的值是 $6.0 + (-1.0)(2) = 4.0$。[6]第二个策略是将 $Z$ 进行变换，使 $Z_t$ 为 0 时，原度量 $Z$ 值正好是有用的那个值，从而得到 $Z_t$，然后再计算相应的回归方程。例如，如果把 2 代入每一个 $Z$ 值中，$Z_t$ 取值范围就变成了

$-1$ 到 3,而 0 对应的就是原来度量中的 2。而在 $Y$ 对 $X$ 和 $Z_t$ 的回归中,得到 $b_1$ 的值为 4.0。后一种策略看上去很繁琐,但作为考察调节变量取某特定值时 $X$ 对 $Y$ 作用大小的方法,后面的章节会说明其优势。

## 显著性检验与置信区间

所有的计算机统计软件的多元回归程序通常都会提供模型估计的回归系数的标准误、显著性检验和置信区间。$b_3$ 的显著性检验的虚无假设是指不存在双线性交互作用。如前所述,$t$ 检验与层级 $F$ 检验产生相同的 $p$ 值,因此,当只在主效应模型中加入一个乘积项时,就不必再进行传统的层级检验。

对交互模型中的“主效应”项,回归系数的条件性本质同样适用于估计系数的标准误。当其他乘积项中的解释变量为 0 时,$X$ 和 $Z$ 在交互模型中所反映的估计标准是条件性的,反映了抽样误差。因此,方程 2.4 中 $b_1$ 的标准误就是当 $Z$ 为 0 时,$Y$ 对 $X$ 回归的系数的抽样误差。简单地讲,方程 2.4 中 $b_2$ 的标准误估计值是当 $X$ 为 0 时,$Y$ 对 $Z$ 回归的系数的抽样误差。

在前面的章节中,我们展示了研究者如何通过方程 2.7 在给定 $Z$ 值时,计算 $Y$ 对 $X$ 回归的斜率。同时,该方程也可计算这一系数的标准误。利用包含乘积项的方程,估计的标准误就是:

$$SE(b_1 \text{ 在 } Z) = [\text{var}(b_1) + Z^2\text{var}(b_3) + 2Z\text{cov}(b_1, b_3)]^{1/2} \qquad [2.9]$$

当 $\text{var}(b_1)$ 是回归系数 $b_1$ 的方差,而 $\text{var}(b_3)$ 是回归系数 $b_3$ 的方差时,$\text{cov}(b_1, b_3)$ 就是回归系数 $b_1$ 和 $b_3$ 的协方差。方程右边的方差和协方差通过标准电脑程序输出得到,尽管默认选项并不输出这些值。给定 $Z$ 值时,$b_1$ 的显著性检验形式是 $t$ 检验,即用系数除以它的标准误,亦即:

$$t = (b_1 \text{ 在 } Z)/SE(b_1 \text{ 在 } Z) \qquad [2.10]$$

此时,$Z$ 是那个检验 $X$ 对 $Y$ 影响时的 $Z$ 值。方程 2.10 的 $t$ 值服从 $t$ 分布,自由度是 $N-k-1$, $k$ 就是在交互作用模型中的解释变量的数目(在此例中, $k=3$)。

一个更简单的策略是使用前面所讨论的变换策略,在任何给定 $Z$ 值的条件

下，计算相关的标准误、置信区间以及 $b_1$ 的显著性检验。通过减去一个常数进行简单的 $Z$ 值变换，使 $Z_t$ 值为 0 对应原先我们感兴趣的 $Z$ 值，然后用电脑程序让 $Y$ 对 $X$、$Z_t$ 和 $XZ_t$ 进行回归。$b_1$ 的取值就是当 $Z_t$ 为 0 时，$X$ 对 $Y$ 作用的系数；$b_1$ 的标准误、显著性检验和置信区间都是此条件性系数所需的参数。下文会展示此方法的应用。

总结一下，传统的交互模型涉及两个连续性解释变量时，是利用下面这个方程对样本数据进行检验的：

$$Y = a + b_1X + b_2Z + b_3XZ + e$$

$b_3$ 在此模型中的系数是单一自由度的交互比较，表示调节变量 $Z$ 变动一个单位，$Y$ 对 $X$ 回归的斜率会变动的单位数量。$b_3$ 的显著性检验的虚无假设是不存在双线性的交互作用。系数 $b_1$ 是当 $Z = 0$ 时，$X$ 对 $Y$ 作用的大小。研究者或许会对给定 $Z$ 值时，$X$ 对 $Y$ 的作用大小感兴趣，此时变换变量策略就有效。这一策略能产生与这些简单效应相关的标准误、显著性检验和置信区间。

## 多重共线性

有些研究者常会害怕包含乘积项的交互分析，因为乘积项与定义这一乘积项的组成变量高度相关。如果 $XZ$ 与 $X$、$Z$ 或两者同时高度相关，那么研究者就会担心多重共线性会使交互作用被低估。这种担忧通常是有误导性的。

如前所述，在对交互作用的显著性进行检验时，$Z$ 减去一个常数，或 $X$ 减去一个常数，或 $X$ 和 $Z$ 都减去一个常数，$b_3$ 的显著性不变。尽管交互作用系数的值和 $t$ 检验都不受变换的影响，但变换确实影响 $XZ$ 和 $Z$ 之间的相关性以及 $XZ$ 和 $X$ 之间的相关性。

变换有时会增加相关性，有时会降低相关性。如果 $X$ 和 $Z$ 都正态分布，那么对中 $X$ 和 $Z$ 后再形成乘积项，会使乘积项与 $X$ 和 $Z$ 都不相关（参见 Cronbach, 1987）。尽管如此，$b_3$ 的显著性检验和 $b_3$ 的置信区间与变换前是相同的，而 $XZ$ 与其组成部分的相关性却相当不同[7]。

在此，并不需要考虑这一结果的统计学原理，感兴趣的研究者可参考

Friedrich(1982)和Cronbach(1987)的著作。高度共线性的主要观点是，乘积项及其组成项之间的共线性并不会使交互分析有问题，除非共线性高到足以破坏电脑代数运算的设计，使统计软件无法分离相关的标准误。如果是这种情况(可从电脑给出的错误提示中得到证实)，研究者可简单地对中 $X$ 和 $Z$(减去各自的均值)，然后再分析。在大部分情况下，这会大大降低共线性，从而解决电脑程序上的问题。

尽管 $XZ$ 和 $X$ 以及 $XZ$ 和 $Z$ 的共线性没有问题，但这并不代表 $X$ 和 $Z$ 的共线性也没有问题。$X$ 和 $Z$ 的高度共线性会导致严重的后果。

## 交互作用的优点

交互作用的优点可通过各种统计方法进行评估，或通过非标准化作用大小的测度，或通过标准化作用大小的测度。最流行的非标准化指标是 $b_3$ 的值。当其他条件不变时，$b_3$ 与 0 的差别越大，说明交互作用越强。最流行的标准化测度是当其组成部分保持不变时，乘积项的半偏相关系数平方。这个值反映了因变量方差的多大比例可由乘积项单独解释，它可通过比较只含主效应模型的复相关系数平方和交互模型的复相关系数平方得到。在避孕措施的例子中，两项加叠模型(只含主效应模型)的复相关系数平方是 0.811，而三项交互模型得到的复相关系数平方是 0.937。因此，交互作用的"解释力"就是 $0.937-0.811=0.126$。这一交互作用在样本中可解释采取避孕措施的意愿的 12.6%的方差。这一指标正向有偏，但当样本规模增大以及 $R^2$ 更高时，这一偏差会相应降低。

对标准化作用大小测度的潜在缺点的讨论，请参见 McClelland 和 Judd(1993)以及 Jaccard(1998)的著作。

## 数字示例

一个具体的例子可以帮助我们总结前面的讨论。一项社会学研究用一个尺度来测量 100 名来自中西部社区天主教徒拥有大家庭的意愿($Y$)。取值范围是 0 到 15，数字越大代表拥有大家庭的意愿更强。另外，受访者被问及在他们的成长过程中，家庭中孩子的个数，以及他们当前的家庭收入，收入以 1000 美元

为单位(即 15 代表 15000 美元)。这 3 个测度的均值和标准差列在表 2.3 中。

**表 2.3 变量均值与标准差**

| 变量名 | 均值 | 标准差 |
|---|---|---|
| 理想的家庭规模 | 4.440 | 2.748 |
| 成长过程中的家庭规模 | 2.960 | 1.601 |
| 家庭收入 | 34.933 | 14.220 |

研究者假设那些成长于大家庭的个人相比那些成长于小家庭的个人,会对拥有大家庭持更正面的态度,因此就有更强的拥有大家庭的意愿。她还假设,这种成长过程中的家庭规模对理想的家庭规模的影响受到家庭收入的调节作用。其逻辑是,个体相对贫穷时,抚养小孩的成本就高,他们就不愿把对大家庭的渴望转化为实际。对富裕家庭而言,这种成本不构成限制条件。因此,研究者希望得到的结果是相比贫穷家庭,在富裕家庭中,成长过程中的家庭规模对拥有大家庭意愿的作用更强。在此研究中,结果变量($Y$)是拥有大家庭的意愿,而某人成长时的家庭规模是关键自变量($X$),收入则是调节变量($Z$)。

在此分析中,$X$ 和 $Z$ 的值都进行了对中处理,以避免多重共线性问题,并使系数 $b_1$ 和 $b_2$ 更易理解。我们计算每一个个案对中后的 $X$ 和 $Z$ 的值,并使用 SPSS 软件进行多元回归分析,用 $Y$ 对 $X$、$Z$ 和 $XZ$ 进行回归(注意,下面 $X$ 和 $Z$ 都指对中转换后的值)。乘积项($XZ$)和其组成项($X$ 和 $Z$)的相关系数非常小($r$ 分别是 $-0.01$ 和 0.019)。

交互模型的复相关系数是 0.725,回归方程是:

$$Y = 4.4279 + 0.81324X + 0.0997Z + 0.0149XZ + e$$

$b_1$、$b_2$ 和 $b_3$ 的标准误分别是 0.098、0.045 和 0.007。

$b_3$ 的 $t$ 检验得到一个统计显著的结果($t = 2.187$, $p < 0.04$),这表明存在交互作用。$b_3$ 的 95%的置信区间是 0.001 到 0.029。半偏相关系数平方表明交互作用的解释力是 0.017,这是将只含主效应模型与交互模型的复相关系数平方相减得到的。对前者而言,复相关系数平方是 0.509;对后者而言,复相关系数平方是 0.526。那么,$0.526 - 0.509 = 0.017$,交互作用可解释理想家庭规模

方差的1.7%。

交互作用的本质通过 $b_3$ 体现。$b_3$ 表明根据收入的不同，成长过程中的家庭规模(RAISE)与理想的家庭规模(FSI)的关系如何变化。家庭收入每增加1000美元(这对应 $Z$ 一个单位)，RAISE 对 FSI 的影响会增加 0.0149 个单位。

为更直观地展示调节变量如何影响 FSI 与 RAISE 的斜率关系，我们在3个不同收入取值上(“低”收入值、“中”收入值和“高”收入值)分别计算用 RAISE 预测 FSI 的简单效应。我们用低于收入均值一个标准差作为“低”收入的标准，而“中”收入指的是收入均值，“高”收入指的是在收入均值一个标准差之上。收入标准差是14.220，那么“低”收入的标准就是 34.933 − 14.220 = 20.713，或20713美元；“中”收入就是34.933或34933美元；“高”收入就是34.933+14.220=49.153，或49153美元。RAISE 对 FSI 影响的斜率正是当收入为“中等”时取得的，因为我们已对数据进行了处理，即对 $X$ 和 $Z$ 都进行了对中处理。$b_1$ 是0.813，$b_1$ 的标准误是0.098，$t$ 值是4.43($p<0.05$)。$b_1$ 的95%的置信区间是0.620到1.009。

大家可利用前面讲解的方法分离“低”和“高”收入时相应的统计量，但我们在此使用的是变换策略。对“低”收入而言，用原收入分数减去20.713，然后用变换后的分数乘以原先的 $X$ 值，再用 $Y$ 对这些相关的解释变量($X$、$Z_t$ 以及 $XZ_t$)进行回归，最后看 $b_1$ 的值。同样，重复这一步骤，计算“高”收入分数($34.933+14.220=49.153$)，然后将原收入分数减去49.153。经过这些步骤得到 $b_1$ 的分析结果如下：

| 收入水平 | $b_1$ | 标准误 | 95%置信区间 | $t$ | $p$ 值 |
|---|---|---|---|---|---|
| 低 | 0.602 | 0.136 | 0.333到0.873 | 4.43 | <0.001 |
| 中 | 0.815 | 0.098 | 0.620到1.009 | 8.33 | <0.001 |
| 高 | 1.026 | 0.136 | 0.757到1.297 | 7.55 | <0.001 |

在“低”收入水平上，成长过程中的家庭每多一个孩子，可使其理想的家庭规模增大0.602；在“中”收入水平上，成长过程中的家庭中每多一个孩子，可使其理想的家庭规模增大0.815；在“高”收入水平上，成长过程中的家庭中每多一

个孩子，可使其理想的家庭规模增大 1.026。成长过程中的家庭规模对理想的家庭规模的影响统计显著，这一点可由 RAISE 作用随收入的增大而增大得到印证，也与研究者假设的 $b_3$ 取值一致。这些简单效应提供给读者关于交互作用的直观感受，但它们不代表对交互作用的正式检验。交互作用的检验基于对 $b_3$ 显著性的检验。

### 图表展示

有些研究者喜欢用图来展示交互作用。方法之一是根据 $Z$ 值的“低”、“中”、“高”分别画 3 条 $X$ 对 $Y$ 作用的直线，如上面计算的那样。相应的斜率分别是 0.602、0.815 和 1.027。要画这种图，我们需要计算每一条回归线的截距。可以利用下面这个公式从最初的回归方程中计算得到：

$$\text{在 } Z \text{ 值上，} X \text{ 对 } Y \text{ 作用的截距} = a + b_2 Z \qquad [2.11]$$

$a$ 指的是交互模型的截距，$b_2$ 是方程中与调节变量相关的回归系数。

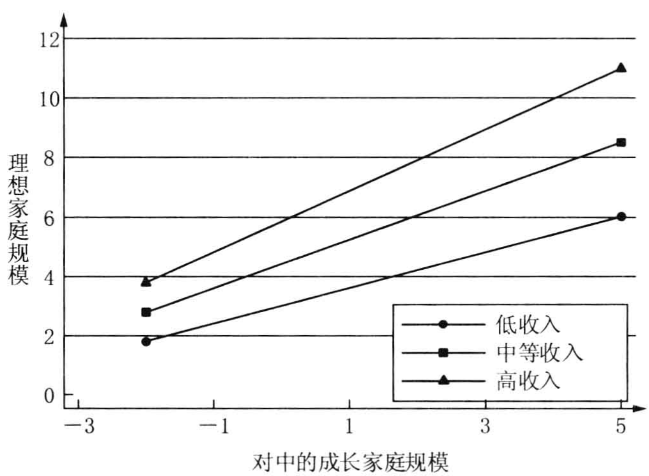



**图 2.1　3 个收入水平成长过程中的家庭规模与理想的家庭规模的回归线**

因为我们使用对中的 $X$ 和 $Z$ 值，因此，“低”收入 $Z$ 是 −14.22，“中”收入是 0，而“高”收入是 14.22。然后，把相应的数字代入方程 2.11，得到：

$$-14.22\text{ 时的截距} = 4.4279 + 0.0997(-14.22) = 3.0102$$

$$0\text{ 时的截距} = 4.4279 + 0.0997(0) = 4.4279$$

$$-14.22\text{ 时的截距} = 4.4279 + 0.0997(14.22) = 5.8456$$

图 2.1 中的回归线确认了 3 个方程 $Y = 3.0102 + 0.603X$、$Y = 4.4279 + 0.815X$ 和 $Y = 5.8456 + 1.026X$ 分别对应“低”、“中”、“高”3 个收入 $Z$ 值。记住，$X$ 在这些方程中都指对中后的 $X$ 值。若无交互作用，那么这 3 条线应当是平行的。很明显，事实并非如此。

## 一个定性解释变量与一个连续解释变量

为了展示一个定性解释变量和一个连续解释变量的情况，我们虚构一个样本。研究者关心青少年对自己与母亲关系的满意程度、该满意程度与性别以及与母亲共处时间之间的关系。

结果测度，即青少年的满意度，是一个从 0 到 21 的尺度，用于测量他们对与母亲的关系的满意程度。分数越高，表明满意程度越高。青少年的性别是虚拟变量，1 = 男性，0 = 女性。母亲与孩子的共处时间是一个母亲自我报告的测度，母亲报告在一个普通的星期中，她与孩子共处的小时数。平均共处时间是 24.06 小时，标准差是 4.62，样本量是 100。首先，我们需要考虑是否可以这样概念化：把共处时间作为关键自变量，把性别作为调节变量，然后修改这两个变量的角色，再重复分析一次。

### 定性调节变量

在此框架下，我们感兴趣的是青少年与母亲的共处时间对男性和女性青少年对其与母亲关系满意度的不同影响。更正式的表述是，当用青少年满意度对共处时间进行回归时，共处时间的系数在男性和女性之间是否存在差异？如果两个斜率相同，那么共处时间对青少年满意度的影响对男性和女性来说就是一样的，并不存在交互作用。然而，若斜率不同，那么性别就调节了共处时间对青

少年关系满意度的影响。

首先，对中时间解释变量，然后，满意度评分($Y$)对对中后的共处时间($X$)、性别虚拟变量($Z$)以及乘积项($XZ$)进行回归。请注意，虚拟变量并未对中。复相关系数的平方为0.784，且统计显著，即 $F(3, 97) = 116.07$，$p < 0.01$。回归方程是：

$$Y = 7.645 + 0.247 Time_c + (-5.270) Gender + (-0.260)(Time_c)(Gender) + e$$

交互作用的回归系数是$-0.260$，其标准误是0.113，95%的置信区间是$-0.484$到$-0.036$。$t$ 检验是统计显著的($t = 2.30$，$p < 0.05$)，表明存在交互作用。乘积项的半偏相关系数平方是0.012，表明交互作用项可解释结果变量样本方差的1.2%。接下来，我们简短地考虑一下对 $b_3$ 的解释。

解释变量 $Time_c$ 的回归系数0.247指的是当 $Gender = 0$ 的时候，共处时间对青少年满意度的作用。因为性别为0对应女性，那么这一系数就是共处时间对女性青少年满意度的影响：与母亲的共处时间每增加一小时，青少年满意度就会增加0.247个单位。这一系数的标准误是0.070(电脑输出结果)，95%的置信区间是0.108到0.386，$t$ 检验统计显著($t = 3.53$，$p < 0.01$)。

上面得到女性共处时间对满意度的影响，而我们还想知道男性的情况。一个简单的方法是对虚拟变量重新编码，使女性=1，男性=0，然后用新的变量值乘以对中的时间变量，再重新计算。回归方程变成：

$$Y = 2.375 + (-0.013) Time_c + (5.270) Gender_r + (0.260)(Time_c)(Gender_r) + e$$

请注意，$b_3$ 的绝对值没变，但正负号反转了。这就是逆向编码的结果，原因在下文会提及。共处时间系数为$-0.013$，这是性别变量逆向编码后，共处时间对青少年满意度的影响。这里，性别为0代表男性，因此是共处时间对男性满意度的影响：与母亲的共处时间每增加一小时，会使男孩的满意度轻微变动，即减少0.013个单位。这一系数的标准误是0.089(电脑输出结果)，95%的置信区间为0.189到0.163，$t$ 检验的结果不显著($t = 0.15$，$p > 0.88$)。

这一关于简单效应的例子表明，青少年满意度对男性共处时间回归的结果

是−0.013,而对女性共处时间回归的结果是 0.247。若计算这两个斜率的差异,得到 −0.013 − 0.247 = −0.260。请注意,这正是最初回归分析中乘积项的系数,即 $b_3$ = −0.260。在传统的乘积项分析中,当乘积项涉及一个连续变量($X$)和一个虚拟变量($Z$)时,乘积项的系数总是等于 $Y$ 对 $X$ 进行两次回归、虚拟变量赋值为 1 的组的(第一次回归)系数减去原参照组重新编码被赋值为 1 的(第二次回归)系数。对 $b_3$ 的显著性检验代表了对上述系数差异的检验,这是单自由度的交互比较。在第一组分析中,$b_3$ 对应 −0.013−0.247 = −0.260;在第二组分析中,性别被重新编码,$b_3$ 对应 0.247 − (−0.013) = 0.260。因此,$b_3$ 的正负号发生变化。

总而言之,使用乘积项回归来计算女性 $Y$ 对 $X$ 回归的斜率,然后对虚拟变量重新编码,再计算男性 $Y$ 对 $X$ 回归的斜率。我们刻画了每一个斜率的显著性检验和置信区间,同时通过乘积项的回归系数($b_3$),我们正式检验了斜率的差异。

我们可以利用调节变量分组,在同一张图中画出不同的回归线。图 2.2 正是这么做的。交互作用表示不平行的两条直线。给定某组回归线的截距,就是当该组为参照组时,整体回归方程的截距。

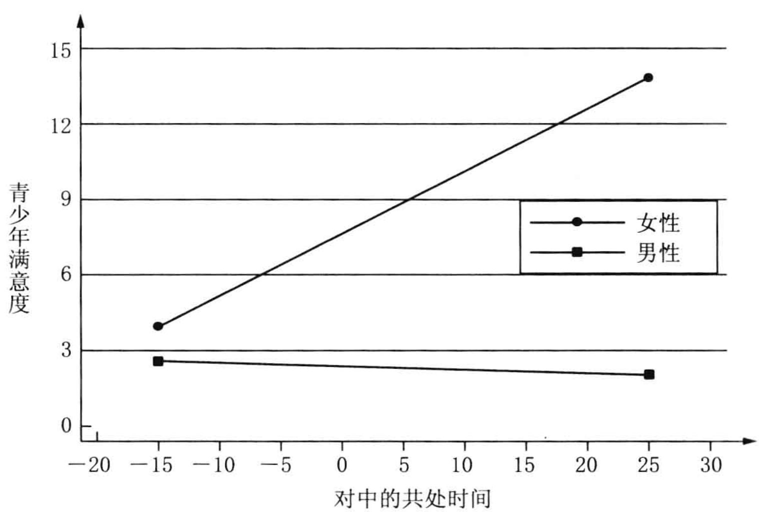



**图 2.2　女性和男性共处时间对关系满意度的影响回归线**

对女性而言，它是 7.645；对男性而言，它是 2.375（请注意，在图 2.2 中，$X=0$ 时的截距出现在图的中央，因为出于美观考虑，我们把 $Y$ 轴左移了）。

在应用研究中，科学家有时会单独计算每一组的回归方程并分别检验 $Y$ 对 $X$ 回归的斜率和显著性，以进行组间比较。当斜率在一组中统计显著，而在另一组中不显著时，就存在组间差异。这种分析往往有很多问题。重要的是，这种分析不能对两组之间的斜率差异进行正式检验，而如果我们要比较组间差异，这种检验就是必须的。一个斜率在一组（的回归）中统计显著，而在另一组（的回归）中统计不显著，但当两组进行正式比较时，斜率差异依然可能不显著。有可能两个斜率都统计显著，但正式的斜率差异统计检验也可能不显著，或者两个斜率都统计不显著，但两者的差异却是统计显著的。研究者需要进行正式的斜率差异比较才能谈论斜率的组间差异，而乘积项策略正是这么做的。另外，分组比较策略的问题是，当进行统计显著性检验时，它忽略了有用的残差信息。在乘积项分析中，给定解释变量上的残差是合并分析的，是根据分析涉及的所有组的信息而给出的。在分组分析中，方差只基于单个组。

## 连续调节变量

假设研究者进行不同的概念化，并调换关键自变量和调节变量的角色。现在考虑的是性别差异对青少年满意度的影响，且这些差异是否为关于与母亲共处时间的一个函数。这一分析使用同一个回归方程，即青少年满意度对对中后的时间变量、性别以及两者乘积项进行回归。我们关心使用不同的概念化和刻画方式时，模型体现出的不同特征。回想一下，回归方程是：

$$Y = 7.645 + 0.247 Time_c + (-5.270) Gender$$
$$+ (-0.260)(Time_c)(Gender) + e$$

与前面一样，$b_3$ 包含交互作用信息，即性别差异如何作为共处时间的一个函数而变动。接下来，我们来看此种理解方式。首先，检查性别的系数。因为性别是一虚拟变量，它的系数代表一个平均的差异，即它是 $Time_c=0$ 时，男性

(虚拟变量中编码为 1 的组)的平均满意度与女性的平均满意度之差。由于 $Time_c$ 是对中后的变量,那么 $Time_c$ 为 0 就对应了原共处时间变量的样本均值(即 24.06 小时)。男性和女性的平均满意度差异(−5.270)是当共处时间为其均值(24.06)时得到的。根据对 $b_2$ 的 $t$ 检验,这一性别差异统计显著($t = 10.28$, $p < 0.01$),而这一差异的标准误是 0.513,95%的置信区间是−6.288 到−4.253。截距项是 7.645,这是当所有解释变量都为 0 时,$Y$ 的平均值。如果 $Time_c$ 为 0,性别为 0,那么方程中所有项都为 0,因此 7.645 就是当 $Time$ 变量等于其样本均值时,女性的平均满意度。如果男性和女性的平均差异为−5.270,而女性的平均值是 7.645,那么简单的代数计算就可告诉我们男性的平均值。它应当是这一差异加上女性的均值,即 $-5.270 + 7.645 = 2.375$。注意,2.375(男性均值)减去 7.645(女性均值)等于平均差异,即−5.270。

若性别差异的估计值对满意度的影响在共处时间的不同取值下结果不同,又会如何?从上面的分析可知,共处时间为 24.06(小时)时,性别差异是 5.270。在其他取值下,情况又如何?为便于理解,让我们计算共处时间为 25.06(小时)时(即共处时间增加一个单位时)的性别差异。我们利用变换策略。首先,把 $Time$ 变量进行变换,使 0 值对应原时间变量的 25.06,即将原时间变量减去 25.06。变换后的时间变量为 $Time_t$,然后乘以虚拟变量性别,这几项同时进入回归方程,得到:

$$Y = 7.892 + 0.247Time_t + (-5.530)Gender$$
$$+ (-0.260)(Time_t)(Gender) + e$$

性别系数变成−5.530,反映的是当 $Time_t$ 为 0 时,平均的性别差异。这意味着,当 $Time = 25.06$ 小时时,性别差异是−5.530。我们将之与先前的分析(共处时间为 24.06 小时,性别差异是−5.270)比较。增加一个时间单位到 25.06小时,这一差异变化了−0.260 个单位,变成了−5.530。现在再看 $b_3$。$b_3$ 等于−0.260,这正是改变调节变量一个单位所带来的平均差异的变化。在传统乘积项分析中,当乘积项涉及一个连续变量($Z$)和一个虚拟变量($X$)时,给定连续变量增加一个单位,乘积项系数反映的是在虚拟变量赋值为 1 的组减去参

照组的平均差异的变化。第二项分析中的截距项 7.892 是当 $Time = 25.06$ 时，女性的平均满意度。同样，可利用代数计算此例中男性的平均满意度，它就是平均差异加上女性的平均满意度，即 $-5.531 + 7.892 = 2.361$。

尽管回归方程包含了所有的相关信息，有些研究者报告的不仅是回归方程及其相关统计量，而且还报告了在 2 个或 3 个给定连续性调节变量取值上，关键自变量的平均差异。这有助于为读者提供直观的对交互作用的感受。例如，我们使用变换策略得到下面这个统计表：

| 共处时间 | 男孩均值 | 女孩均值 | 平均差异 | 95%置信区间 | $t$ | $p$ 值 |
|---|---|---|---|---|---|---|
| 20 小时/周 | 2.43 | 6.64 | −4.21 | −5.54 到−2.98 | 4.43 | <0.001 |
| 25 小时/周 | 2.36 | 7.88 | −5.52 | −6.59 到−4.44 | 8.33 | <0.001 |
| 30 小时/周 | 2.30 | 9.11 | −6.82 | −8.63 到−5.00 | 7.55 | <0.001 |

当每周共处时间为 20 小时时，性别差异对关系满意度的影响是 −4.21(即女性比男性的满意度高 4.21 个单位)；当每周共处时间增加到 25 个小时时，性别差异变成 −5.52；若每周共处 30 个小时，差异增加更多，达到 −6.82。这一展示性统计提供了直观的交互作用效应，我们看到，均值差异随着调节变量取值的不同而变动。然而，应当谨记，这些简单效应不是交互作用检验的核心。交互作用的正式检验应当在 $b_3$ 中。

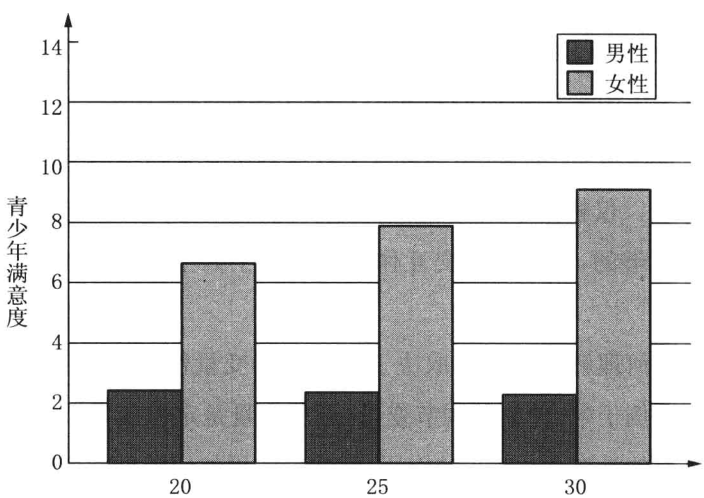



**图 2.3　选定共处时间点上不同性别间关系满意度差异的均值**

这一结果可用直方图展示，如图 2.3。若不存在交互作用，那么每一个共处时间点上男性和女性柱高度的差异就应相同。缺乏这种一致性就表明存在交互作用。该图表明，当与孩子共处的时间增加时，性别差异对关系满意度的影响变大了。另外一个图形展示方法是画出由关键自变量定义的不同组之间，$Y$ 对 $Z$ 的单独的回归线。这已由图 2.2 进行了展示。两条回归线之间的距离在水平线（调节变量）上的差异反映了 $Y$ 在这一点上的平均差异。图 2.2 包含的信息比图 2.3 更多，但也有些研究者认为图 2.3 更有意思。

### 两组以上的定性变量

有一些应用涉及定性和连续性解释变量，且定性变量包含两个以上的层级。上面这些一般化的原则都适用于这种情况。

请考虑这样的情况：结果变量是收入（$Y$），两个解释变量是受教育年限和种族。假设种族有 3 类：美国黑人、美国白人[8]、拉美人。这由两个虚拟变量表示。假设第一个虚拟变量是 $D_{AA}$，所有美国黑人赋值为 1，而其他人赋值为 0。第二个虚拟变量是 $D_L$，所有拉美人赋值为 1，其他人赋值为 0。以美国白人为参照组。为了表示两个解释变量的交互作用，我们构建代表一个解释因素的变量和代表另一个解释因素变量之间所有可能的乘积项。假设 $ED$ 代表教育测度，我们就得到乘积项（$ED$）（$D_{AA}$）和（$ED$）（$D_L$）。教育和种族整体的交互作用要求使用方程 1.1 中的层级回归检验。只含主效应模型中的复相关系数平方（只含 $ED$、$D_{AA}$ 和 $D_L$ 的模型）与包含主效应和所有乘积项模型的复相关系数平方进行比较。仅根据乘积项单个 $b$ 系数的显著性检验来检查整体交互作用的显著性是不够的，因为整体交互作用具有一个以上的自由度，必须使用层级策略检验。

研究者如何理解回归方程取决于挑选哪个变量作为调节变量这一问题？假设在第一个例子中，种族是调节变量（调节变量是定性变量）。在此情况下，研究者感兴趣的是教育对收入的作用，且这种作用如何随种族变化而变化。回归方程的一般形式是：

$$Y = a + b_1 ED + b_2 D_{AA} + b_3 D_L + b_4 ED(D_{AA}) + b_5 ED(D_L) + e \qquad [2.12]$$

正如前面所讨论的,乘积项回归系数反映了斜率的差异。$b_4$ 关注 $Y$ 对 $ED$ 回归的斜率,并估计了 $D_{AA}$ 为 1 的组的斜率减去参照组斜率的差异。换言之,它反映了美国黑人与美国白人的斜率差异。$b_5$ 对应同样的斜率差异,指拉美人和美国白人的斜率差异。显著性检验、标准误、置信区间提供单自由度交互比较通常所关心的额外信息。

那么,如果除了由 $b_4$ 和 $b_5$ 反映的这两个比较外,我们还想评估 $Y$ 对 $ED$ 回归时,美国黑人和拉美人的斜率差异呢?这种比较和它的相关显著性检验可以通过产生一组新的虚拟变量和乘积项,或把参照组改成美国黑人或拉美人,然后在新回归方程中检查乘积项系数,就能分离出来了。第 4 章会讨论如何进行这种多元比较,并调整多重比较带来的扩大的误差率。

在方程 2.12 中,$ED$ 的系数 $b_1$ 是一个简单效应,它指的是当所有种族虚拟变量都为 0 时,教育对收入的作用。$D_{AA} = 0$ 且 $D_L = 0$ 指参照组,因此 $b_1$ 是在美国白人中,教育对收入的作用。标准误、置信区间和显著性检验都针对这一作用,可从 $b_1$ 的电脑输出结果中看到。如果需要分离出拉美人相应的简单效应,那就需要重新定义一组新的虚拟变量和乘积项,使得拉美人成为参照组,再重新进行电脑分析。此过程与美国黑人的分析过程相似。

请考虑这样的情况:教育是调节变量。现在我们感兴趣的是种族差异对收入的影响,且这些差异如何受到教育的影响。使用同样的回归方程,但研究者在使用这种方法之前,通常会对中教育(或者把教育对中在一个有意义的值上)。下面讨论中的对中操作均如此。

首先,我们考虑 $D_{AA}$ 的系数,即 $b_2$。这一系数是当 $ED_c = 0$ 时,$D_{AA}$ 这一虚拟变量取值为 1 的组和参照组的平均收入差异,这个差异就是当教育等于样本均值时,美国白人和美国黑人之间平均的收入差异。$D_{AA}$ 和 $ED_c$ 的乘积项系数 $b_4$ 反映了教育增加一个单位所导致的平均收入差异的变化。$D_L$ 的系数是 $b_3$,代表当教育等于样本均值时,拉美人和美国白人直接的平均收入差异。$D_L$ 和

$ED_c$ 的乘积项系数 $b_5$ 反映了教育增加一个单位时,拉美人和白人的平均收入差异的变动情况。如果研究者对拉美人和美国黑人可比的系数和显著性检验感兴趣,那么只需重新定义虚拟变量,使美国黑人或拉美人成为参照组,然后再进行电脑分析。使用上述讨论的原则就可分离出相应的系数。

回归方程中的截距是当 $ED_c = 0$ 时,参照组(美国白人)的平均收入。

## 交互的形式

**表 2.4　两种交互作用**

| $Z$ 值 | 例 A:两组之间平均差异 | 例 B:两组之间平均差异 |
|---|---|---|
| 1 | 2.0 | 2.0 |
| 2 | 4.0 | 4.0 |
| 3 | 6.0 | 6.0 |
| 4 | 8.0 | 8.0 |
| 5 | 10.0 | 10.0 |
| 6 | 12.0 | 10.0 |
| 7 | 14.0 | 10.0 |

如前所述,当乘积项中两个变量都是连续变量时,传统交互模型评估的是一种特殊的交互形式,即双线性交互。当乘积项中的一个是定性变量,另一个是连续变量时,传统交互模型估计的是一种特定的交互。这个模型可通过两种不同的情况来说明,尽管两种情况反映了同一个模型。一种情况是调节变量为定性变量。在这种情况下,预先假定在调节变量的每个类别上,$Y$ 和 $X$(结果变量和连续性关键自变量)的直接关系为线性关系都为真。如果至少在调节变量的某一个类别上,$Y$ 和 $X$ 的关系是非线性的,那么传统乘积项模型的设定就是错的,此时就需要另一模型(见第 4 章的讨论)。另一种情况是调节变量为连续变量。在此情况下,预先假定组均值的有序变化是 $Z$ 的线性函数。请考虑表 2.4中的两个例子。例 A 表示的是两组之间的差异是 $Z$ 的线性函数:$Z$ 每增加一个单位,组间的平均差异就增加两个单位。而例 B 则反映了非线性关系不能用传统的乘积项方法来建模。在 $Z$ 的低层级上,$Z$ 每增加一个单位,组间的平均差异就增加两个单位,但这最终渐近并平稳下来(不再增加)。

## 小结

研究者经常关注双向交互的交互模型。交互中所涉及的两个变量本质上都可以是连续的,或一个连续、一个定性。在这两种情况下,研究者都宣称一个是关键变量,另一个是调节变量。恰当的乘积项可以通过计算得到,然后交互模型就包括了"主效应"和乘积项。当整体交互作用是单自由度时,整体显著性检验就反映在乘积项的回归系数上。当整体交互有一个以上的自由度时,就需要运用层级检验比较只含有主效应模型和交互模型的差异来估计整体交互作用的统计显著性。

非标准化系数的交互效应的优势在于,其回归系数反映了单自由度交互作用。标准化回归系数的优势在于,交互项的半偏相关系数平方反映了交互作用的大小。

典型的对交互作用的理解关注的是与关键自变量相关的回归系数以及与乘积项相联系的回归系数,这取决于哪个变量是调节变量。不同的对中策略和重新编码方法可帮助读者理解交互作用的本质。

# 第 3 章 | 三向交互作用

本章把第 2 章的原则扩展到三向交互作用分析上。首先考察全部是连续变量的情况，然后考察定性和连续变量混合的情况。

## 连续解释变量

假定有一项研究，一名行为医学研究者关注父母为了预防某种形式的肝炎，给孩子注射疫苗的意愿。这种意愿测度的取值是 0 到 100，分数越高说明接种疫苗的意愿越强烈。研究者研究 3 个影响接种疫苗意愿的潜在变量。第一个变量是如果父母不给孩子接种疫苗，预计孩子感染这种疾病的可能性。我们称这个变量为疾病的“预期易染性”，用 0 到 100 的分数进行测量，分数越高，疾病传染性越强。第二个变量是如果孩子真的感染了某种肝炎，那么父母预计这种病会有多严重的后果。这一变量被称为“预期严重性”，也用 0 到 100 分测量，分数越高，说明预期的严重性越强。第三个变量是预计接种疫苗使孩子避免染病的可能性。这个变量又被称为“预期有效性”，用 0 到 100 分测量，分数越高，说明预期有效性越强。接种疫苗的意愿受到这 3 个变量交互函数的影响。此研究的样本是 200 名家长。

为分析三向交互作用，我们需要确定一个关键自变量和两个调节自变量来明确交互的逻辑。对三向交互作用而言，需要区分两个调节变量。假设研究者选择的关键自变量是预期易染性，并假设接种疫苗的意愿是预期易染性的函数。具体而言，假设预期易染性越高，接种的意愿也会越强。进一步的假设是，预期易染性对接种意愿的作用会受到预期严重性的调节作用。如果父母认为

这个疾病并不严重，那孩子易受传染这一问题就无关紧要了。父母缺乏动力去为他们认为无关痛痒的疾病寻求疫苗。然而，当预期严重性增强时，预期易染性就会影响接种疫苗的意愿。在此情况下，预期严重性被称为“一阶调节变量”，因为它直接调节了预期易染性对接种疫苗意愿的影响。研究者进一步假设，预期严重性的调节作用会受到预期有效性的影响。如果父母认为疫苗不那么有效，那么无论多么容易感染这种疾病或者万一染病，后果有多么严重，他们都不会要求接种。然而，当疫苗的预期有效性增强时，预期严重性和预期易染性的交互机制会自我增强。疫苗的预期有效性是二阶调节变量，因为它调节了一阶调节变量对关键自变量和因变量的作用。当然，这并非三向交互作用必须的概念化方式。然而，我们发现，一阶和二阶调节变量的理解方式对思考三向交互作用很有用。并且，当调查者描述三向交互作用时，他们总是采用这种方式(尽管有时候不那么明显)，从而使复杂的关系变得容易理解。

为了分析传统的交互模型，需要形成乘积项并加入主效应模型反映双向交互和三向交互。主效应模型是：

$$\dot{Y} = a + b_1X + b_2Z + b_3Q + e$$

在这里，$X$ 是预期易染性，$Z$ 是预期严重性，$Q$ 是预期有效性。三向交互作用模型加入了所有这 3 个解释变量成对的项($XZ$、$XQ$ 和 $ZQ$)以及 3 个解释变量共同的乘积项($XQZ$)。模型就变成如下形式：

$$Y = a + b_1X + b_2Z + b_3Q + b_4XZ + b_5XQ + b_6QZ + b_7XZQ + e \quad [3.1]$$

三向交互作用的显著性就是 $b_7$ 的显著性。用标准术语来讲，三向交互作用的解释力是用方程 3.1 的复相关系数平方减去一个基于方程 3.1 但省略 $XZQ$ 这一项的方程的复相关系数平方。低阶系数的理解总是条件性地取决于高阶乘积项，而条件化是指其他变量在高阶交互项上等于 0。例如，$b_4$ 代表当 $Q=0$ 时，$X$ 和 $Z$ 对 $Y$ 的双向交互作用；$b_1$ 代表当 $Z=0$ 且 $Q=0$ 时，$X$ 对 $Y$ 的作用。关键自变量、一阶调节变量、二阶调节变量的区分有助于我们理解方程的系数。

回到例子中，先对中所有解释变量，然后再形成相关乘积项。SPSS 回归输出结果列在表 3.1 中。三向乘积项的系数是 0.0008137，且统计显著($t=$

4.715，$p < 0.01$，95%的置信区间是0.000473到0.001154)，表明存在三向交互作用。而省略三向乘积项方程的复相关系数平方是0.400，完整模型的复相关系数平方是0.462。那么这两个复相关系数平方的差异为0.062，表明三向交互作用能够解释接种疫苗意愿的方差的6.2%。

基于前面展示的原则，预期易染性($X$)系数是0.219，这代表预期严重性($Z$)和预期有效性($Q$)都为0(即当预期严重性和预期有效性为"平均"或"中等"，因为对它们进行过对中处理)时，预期易染性对接种意愿的影响。0.219表明预期易染性每增加一个单位，接种意愿就增加0.219个单位($t = 5.062$，$p < 0.01$，95%的置信区间是0.134到0.304)。而$XZ$的系数反映了当二阶调节变量预期有效性为0(即预期有效性是"中等"或"平均")时，预期易染性与一阶调节变量预期有效性的交互作用。这一系数是0.01227($t = 4.61$，$p < 0.01$，95%的置信区间是0.007到0.018)，这是预期有效性为"平均值"时，预期易染性每增加一个单位，预期易染性对接种意愿斜率的增加值。控制预期有效性为其样本均值时，预期严重性每增加一个单位，预期易染性对接种意愿的影响力就增加0.01227个单位。这一系数的理解与第2章中的双向交互作用一样，但它是以$Q = 0$为条件的(在此例中，代表对中后的预期有效性为0)。

为理解三向乘积项系数的意义，我们要重新计算当预期有效性($Q$)比其均值高一个单位而非其均值时，此例中$XZ$的系数。将预期有效性减去一个常数45.79，而非减去样本均值44.79即可。这样，我们把变换后的预期有效性0值定义为45.79，而非原尺度中的44.79。然后，使用变换后的分数重新计算乘积项，并重新进行回归。

正如第2章讨论的，复相关系数平方和$b_7$的值在这个变换之后保持不变。在新分析中，$XZ$乘积项系数为0.0130837。通过计算前后两次回归中$b_4$的差异来看双向交互系数变化，即$0.0130837 - 0.01227 = 0.0008137$。检查表3.1中的三向交互作用系数，正好是0.0008137，正是预期有效性变动一个单位时，双向交互作用变动的量。对一个由3个连续变量组成的三向交互模型而言，关于$X$、$Z$和$Q$以及它们之间不同的乘积项，我们令$X$是关键自变量，$Z$是一阶调节变量，而$Q$是二阶调节变量。三向交互作用的系数是给定$Q$变动一个单位，

**表 3.1 SPSS 三向交互作用输出结果**

| 模　型 | $R$ | $R^2$ | 调节的 $R^2$ | 测量的 SE |
|---|---|---|---|---|
| 1 | 0.680 | 0.462 | 0.443 | 9.85080 |

| 模　型 | 非标准化回归系数 | | 标准化回归系数 | $t$ | Sig. | B 的 95%置信区间 | |
|---|---|---|---|---|---|---|---|
| | B | SE | β | | | 低　端 | 高　端 |
| （常数项） | 33.332 | 0.697 | | 47.849 | 0.000 | 31.958 | 34.706 |
| 易染性($X$) | 0.219 | 0.043 | 0.270 | 5.062 | 0.000 | 0.134 | 0.304 |
| 严重性($Z$) | 0.203 | 0.044 | 0.245 | 4.629 | 0.000 | 0.116 | 0.289 |
| 有效性($Q$) | 0.248 | 0.044 | 0.299 | 5.635 | 0.000 | 0.161 | 0.334 |
| $XZ$ | 0.01227 | 0.003 | 0.246 | 4.606 | 0.000 | 0.007 | 0.018 |
| $XQ$ | 0.01254 | 0.003 | 0.243 | 4.519 | 0.000 | 0.007 | 0.018 |
| $QZ$ | 0.01383 | 0.003 | 0.263 | 4.951 | 0.000 | 0.008 | 0.019 |
| $XQZ$ | 0.0008137 | 0.000 | 0.254 | 4.715 | 0.000 | 0.000 | 0.001 |

$X$ 和 $Z$ 双向交互作用系数的变动数量。如果重新回归，使预期有效性($Q$)高于其样本均值两个单位进行对中处理，那么 $XZ$ 的系数就变成：

$$0.01227 + 0.0008137 + 0.0008137 = 0.0138974$$

尽管回归方程中包含了所有这些理解三向交互作用所需的相关信息，但大部分读者还是感到从方程本身获取三向交互作用的直观理解依然有些困难。正如在第 2 章中我们发现，提供不同的 $Z$ 和 $Q$ 取值组合时，$X$ 对 $Y$ 作用的斜率有助于读者思考交互作用。其结果列在表 3.2 中。此表有 4 种情况，在这些情况下观察 $Y$ 对 $X$ 回归的斜率：(1)$Z$ 值低，$Q$ 值低；(2)$Z$ 值低，$Q$ 值高；(3)$Z$ 值高，$Q$ 值低；(4)$Z$ 值高，$Q$ 值高。低和高都被定义为低于和高于 $Q$ 和 $Z$ 各自样本均值的一个标准差。$Y$ 对 $X$ 在每种情况下的回归斜率展示在一个 $2\times2$ 的析因分析表中，行的值代表一阶调节变量的低和高，而列的值代表二阶调节变量的低和高。这些系数的值的显著性检验及其置信区间都通过计算机计算得到，并通过变换策略分离出相关系数，观察其对预期易染性的影响。此表说明，预期有效性相对较低时，预期易染性对接种疫苗意愿作用的影响可忽略不计，且相比预期有效性相对高时，意愿要小得多。事实上，只有当预期严重性和预期有效性都很高时，才能看到预期易染性对接种意愿有实际意义的影响力。

**表 3.2　作为预期严重性和预期有效性函数的接种意愿对预期易染性回归的斜率**

| | 低有效性 | | | 高有效性 | | |
|---|---|---|---|---|---|---|
| | 斜率 | 95%置信区间 | $t$ | 斜率 | 95%置信区间 | $t$ |
| 低严重性 | 0.003 | −0.128 到−0.188 | 0.376 | 0.009 | −0.183 到 0.200 | 0.089 |
| 高严重性 | 0.016 | −0.145 到 0.178 | 0.197 | 0.820 | 0.648 到 0.993 | 9.39* |

注：* 表示 $p<0.01$。

这个统计结果可用一个并排图形来展示。一张图是当 $Q$ 为“低”时，通过 $Z$ 的两个值得到的回归线；另一张图是 $Q$ 为高时，通过 $Z$ 的两个值得到的回归线。这两张图就可并排展示。图 3.1 正是采用这种展示方法。若不存在三向交互作用，那么这两个斜率在一张图上将是平行的，而在另一张图上也应如此。但我们的数据表明并非如此。

为构建此图，我们需要计算给定的 $Z$ 值和 $Q$ 值时，$Y$ 对 $X$ 回归的截距。截距就是在完整的交互回归中，使用变换后的 $Z$ 值和 $Q$ 值分离出感兴趣斜率的截距项。

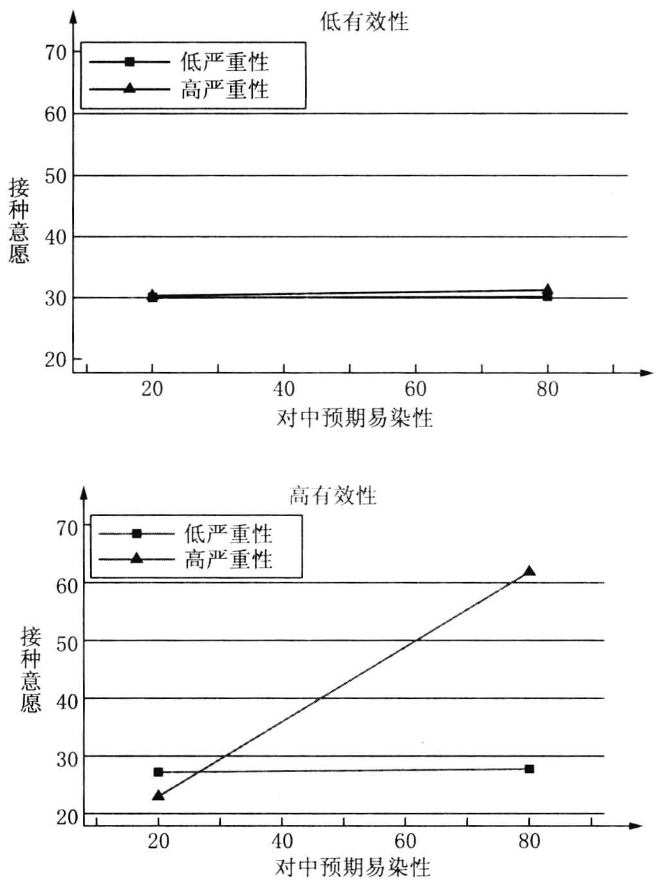



**图 3.1　作为预期严重性和预期有效性函数的预期易染性对接种意愿的作用的回归线**

## 定性和连续解释变量

假设一位研究者对司法系统中的偏见感兴趣。他设计了一个研究，使参与者阅读一段针对被告所作出的法庭案例分析报告，然后给出此被告被判定为有罪的可能性分数，分数由 0 到 100。结果变量从 0 到 100，分数越高表示越高的

有罪概率或越高的被判定有罪的可能性。所有参与者都阅读相同的报告，只不过其中一半人被告知被告是美国黑人，而另一半人则被告知被告是美国白人。参与者中有一半是美国黑人，而另一半是美国白人。这形成一个 $2\times2$ 的析因设计，交叉的是被告的种族和“陪审员”（研究参与者）的种族。研究者同时测量了第三个变量，即参与者具有自由主义或保守主义倾向，相关尺度是从 0 到 100 的分数，分数越高表示越强的自由主义倾向。被告的种族由一个虚拟变量 $AC$ 表示，美国白人是参照组。陪审员种族同样由一个虚拟变量表示——$JU$，美国白人是参照组。自由主义倾向尺度 $L$ 进行对中操作，由此产生相应的乘积项（因此，此处 $L$ 代表对中后的自由主义—保守主义尺度）。表 3.3 是 SPSS 回归程序的输出结果。

三向乘积项的系数 1.007 统计显著（$t=7.31$，$p<0.01$，95% 的置信区间为 0.735 到 1.279），表明存在三向交互作用。省略 $AC\times JU\times L$ 模型的复相关系数平方是 0.595，全模型的复相关系数平方是 0.782。两者之差为0.192，表明三向交互作用可以解释 19.2%的定罪方差。接下来，我们在两种情况下讨论这些系数：一是关键自变量为连续变量；二是关键自变量为定性变量。

## 连续关键自变量

在此框架下，研究者感兴趣的是自由主义—保守主义倾向对定罪的影响。研究者指出，更具自由主义倾向的陪审员相比保守的陪审员，较不倾向于判定被告有罪。然而，这种自由主义—保守主义倾向对定罪的影响会受到被告种族的调节作用。当被告是美国黑人时，相比被告是美国白人，自由主义—保守主义倾向将会是一个对定罪更有力的解释变量。因此，被告种族就是一阶调节变量。根据三向交互作用原理，被告种族对自由主义—保守主义倾向对定罪的调节作用还取决于陪审员自身的种族。研究者预测，前述（被告）种族和（陪审员）自由主义—保守主义分数的双向交互作用的机制在美国白人陪审员中（相比美国黑人陪审员）更为显著。因为在司法案件中，相比美国黑人陪审员，偏见和种族被认为在美国白人陪审员身上体现得更为明显。陪审员种族因此就是一个二阶调节变量。

表 3.3 陪审员例子中的 SPSS 输出

| 模型 | $R$ | $R^2$ | 调节的 $R^2$ | 测量的 SE |
|---|---|---|---|---|
| 1 | 0.884 | 0.782 | 0.775 | 8.58238 |

| 模型 | 非标准化回归系数 | | $t$ | Sig. | B 的 95%置信区间 | |
|---|---|---|---|---|---|---|
| | B | SE | | | 低端 | 高端 |
| (常数项) | 39.006 | 1.111 | 35.102 | 0.000 | 36.817 | 41.196 |
| 被告种族($AC$) | 17.235 | 1.570 | 10.980 | 0.000 | 14.142 | 20.328 |
| 陪审员种族($JU$) | 1.228 | 1.570 | 0.782 | 0.435 | −1.865 | 4.320 |
| 自由主义($L$) | −0.576 | 0.068 | −8.473 | 0.000 | −0.710 | −0.442 |
| $AC*JU$ | −17.241 | 2.218 | −7.772 | 0.000 | −21.611 | −12.870 |
| $AC*L$ | −0.952 | 0.098 | −9.717 | 0.000 | −1.145 | −0.759 |
| $JU*L$ | −0.033 | 0.098 | −0.340 | 0.734 | −0.226 | 0.160 |
| $AC*JU*L$ | 1.007 | 0.138 | 7.307 | 0.000 | 0.735 | 1.279 |

自由主义—保守主义系数在表 3.3 中是 −0.576，反映的是当 $AC = 0$ 且 $JU = 0$ 时，自由主义—保守主义倾向对定罪的影响。这表示当美国白人被告受到美国白人陪审员审判时的情况，它是简单效应。如果先分离出其他 3 种情况下的简单效应，可使方程中其他系数的含义更明确。这可通过为 $AC$ 和 $JU$ 重新计算一组新的虚拟变量而达成，即重新定义参照组，从而使与自由主义—保守主义相关的系数代表我们感兴趣的简单效应。例如，重新定义 $AC$ 的参照组为美国白人，而 $JU$ 的参照组为美国黑人，那么重新计算乘积项并重新进行回归，得到与 $L$ 相连的系数就是美国白人被告被美国黑人评判时，自由主义—保守主义倾向对定罪的影响。表 3.4 展示了 4 种不同情况下的电脑分析结果。

**表 3.4 自由主义对定罪的作用斜率**

| | 斜率 | 标准误 | 95%置信区间 | $t$ | $p$ 值 |
|---|---|---|---|---|---|
| AA 陪审员，AA 被告 | −0.554 | 0.066 | −0.685 到 −0.424 | 8.35 | <0.001 |
| AA 陪审员，EA 被告 | −0.609 | 0.071 | −0.748 到 −0.470 | 8.64 | <0.001 |
| EA 陪审员，AA 被告 | −1.528 | 0.071 | −1.667 到 −1.389 | 21.65 | <0.001 |
| EA 陪审员，EA 被告 | −0.576 | 0.068 | −0.710 到 −0.442 | 8.473 | <0.001 |

注：AA = 美国黑人；EA = 美国白人。

仔细检查表 3.3 中关键自变量与一阶调节变量乘积项（$AC \times L$）的系数。这一系数反映了一个定类变量和一个连续变量之间的双向交互作用，所以它反映了斜率差异。具体而言，它指 $JU = 0$，定罪分数对自由主义—保守主义倾向回归时，黑人被告与白人被告的自由主义变量的斜率差异。$JU$ 为 0 代表陪审员为白人，因此只关注白人陪审员的情况。这一系数等于 −0.952。查阅表 3.4 并只观察白人陪审员，我们发现自由主义—保守主义对黑人被告的归罪贡献系数是 −1.525，而对白人被告是 −0.576。两个斜率之差是 (−1.528) − (−0.576) = −0.952，这正是双向交互作用参数。

从表 3.3 可见，这一斜率差异统计显著（$t = 9.72$，$p < 0.01$，95% 的置信区间为 −1.145 到 −0.759）。对包含一个连续解释变量 $X$、两个定性解释变量 $Z$ 和 $Q$ 以及它们相应乘积项的交互模型来说，令 $X$ 为关键自变量，$Z$ 为一阶调节变量，而 $Q$ 为二阶调节变量。对用虚拟编码方法的定性变量而言，乘积项 $XZ$

的回归系数就是斜率之间的差异。它关注虚拟变量 $Z$ 赋值为 1 的组与参照组之间的 $X$ 对 $Y$ 作用斜率的差异，当然，这些都只针对 $Q$ 为参照组的情况。

对白人陪审员而言，黑人被告和白人被告之间的斜率差异是－0.952，那么黑人陪审员的情况又如何呢？这个差异的值、标准误、置信区间都可通过重新定义 $JU$ 这一虚拟变量，使美国黑人陪审员成为参照组而实现。再重新计算乘积项，重新进行回归，检查 $AC \times L$ 乘积项系数。在新的分析中，这个系数是 0.055。这确认了表 3.4 中的相关的斜率是－0.554 和－0.609，他们的差异是 $(-0.554)-(-0.609)=0.055$。这个差异统计不显著（$t=0.565$，$p<0.57$，95% 的置信区间为－0.136 到 0.246）。

总的来说，对美国黑人陪审员而言，黑人被告 $Y$ 对 $X$ 的斜率减去白人被告 $Y$ 对 $X$ 的斜率，其差是 0.055，这可由美国黑人陪审员（与自由主义分数 $L$）的双向交互参数得到。美国白人陪审员的斜率差异是－0.952，它代表美国白人陪审员（与自由主义分数 $L$）的双向交互作用参数。如果不存在三向交互作用，那么这些双向交互参数应当相同（假设不存在样本偏差）。事实上，它们确实提示可能存在三向交互作用。[9] 两个双向交互参数的差异是 $(0.055)-(-0.952)=1.007$。再看三向乘积项的系数，注意它等于 1.007。而这一参数的显著性检验评估的是在假设双向交互作用无差异这一虚无假设为真时，观察到结果为 1.007或者更大结果的概率。对有一个连续变量 $X$、两个定性变量 $Z$ 和 $Q$ 及其相应乘积项的交互模型，令 $X$ 作为关键自变量，$Z$ 为一阶调节变量，$Q$ 为二阶调节变量。对虚拟编码的定性变量而言，三向乘积项的回归系数指的是斜率差异的不同。差异是当 $Z$ 赋值为 1 的组的 $Y$ 对 $X$ 的斜率减去参照组的 $Y$ 对 $X$ 的斜率。这一斜率差异对虚拟变量 $Q$ 而言，即当 $Q$ 赋值为 1 的组的 $Y$ 对 $X$ 的斜率值减去 $Q$ 为参照组的 $Y$ 对 $X$ 的斜率。

简言之，表 3.4 提供了不同的实验情境下 $Y$ 对 $X$ 的斜率，并确认了这些简单效应的显著性检验。三向交互作用隐含了双向交互作用会根据二阶调节变量取值的变化而变化。表 3.4 中斜率的模式很明显地展示了三向交互作用的来源。白人陪审员的斜率差异比黑人陪审员的斜率差异大得多。对应由 $Z$ 定义的两组，在每个 $Q$ 的取值上单独作图，进而得到的表示 $Y$ 对 $X$ 的回归线的并排

图可展示这一点。

## 定性关键自变量

我们可使用定性解释变量作为关键自变量，重新分析上面的研究。概念化的逻辑是这样的：研究者预计由于负面的刻板印象和偏见，美国黑人被告会比美国白人被告更容易被判有罪。然而，这种倾向会受到陪审员种族的考验（一阶调节变量）。只有当陪审员是美国白人时，才会出现美国黑人被告比白人被告更易被判有罪的情况。这一假设源于白人陪审员会表示出偏见，而黑人陪审员则不会。最后，陪审员种族的调节效应依赖于陪审员本身的自由主义—保守主义倾向。对那些自由主义的陪审员而言，这不是问题，即无论陪审员是黑人还是白人，都不会对黑人被告有偏见。

三向交互作用的检验与前文中的完全一样，我们这里仍用表 3.3 的结果。其焦点是使用上述概念框架，在不同的低阶系数上展示三向交互作用。

确认三向交互作用存在时，首先检查与关键自变量和一阶调节变量联系在一起的系数，即乘积项 $AC \times JU$。这一系数是 $-17.241$。如果乘积项中的两个自变量都是虚拟变量，那么这些相关系数可反映在一个 $2 \times 2$ 的表中。

**表 3.5　自由主义倾向分数取均值、定罪概率作为被告种族和陪审员种族的函数**

| | 均　值 | 标准误 | 95%置信区间 |
|---|---|---|---|
| AA 陪审员，AA 被告 | 40.23 | 1.11 | 38.04 到 42.41 |
| AA 陪审员，EA 被告 | 40.23 | 1.11 | 38.04 到 42.41 |
| EA 陪审员，AA 被告 | 56.24 | 1.11 | 54.06 到 58.43 |
| EA 陪审员，EA 被告 | 39.00 | 1.11 | 36.62 到 41.20 |

注：AA = 美国黑人；EA = 美国白人。

这个 $2 \times 2$ 表格交叉的是第一个虚拟变量取值为 1 时相比其参照组，以及第二个虚拟变量取值为 1 时相比其参照组。表 3.5 识别出这一子表。我们可识别每个单元格的平均被判有罪概率的均值，当然，条件是对中的自由主义—保守主义分数为 0。全回归模型中的截距是当 $AC = 0$、$JU = 0$ 且 $L = 0$ 时，截距的有罪判定的均值。因此，当自由主义—保守主义分数是平均值时，美国白人

被告由美国白人陪审员来评分的平均有罪判定分数是 39.006(95%的置信区间是 36.817 到 41.196)。可通过代数(如第 4 章讨论的)或通过重新编码虚拟变量并重新进行回归,从而得到其他 3 个单元格的均值。在后一个方法中,我们重新定义虚拟变量从而使参照组对应 2×2 表格中我们感兴趣的单元格,然后重新计算乘积项,重新进行回归,并注意新方程中截距项的值。使用这一策略,就会得到表 3.5 中 4 个单元格的均值及其置信区间。

如第 1 章所讨论的,2×2 表中交互作用的参数是调节变量在第一个类别上关键自变量取值的差异减去调节变量在第二个类别上关键自变量取值的差异,即 (40.23－40.23)－(56.24－39.00)＝－17.24。请注意,这个值等于 $AC \times JU$ 的回归交互作用。标准误、置信区间和显著性检验都提供了检验 2×2 表格中交互作用有效性的视角。

对交互模型而言,有两个定性解释变量 $X$ 和 $Z$、一个连续性解释变量 $Q$ 以及它们之间产生的乘积项。令 $X$ 为关键自变量,$Z$ 为一阶调节变量,$Q$ 为二阶调节变量。对虚拟编码的定性变量而言,$XZ$ 的回归系数代表两个均值的差异。它是当 $Q = 0$(对中后 $Q$)时,$X$ 取值为 1 的组的均值减去参照组的均值之差,再减去 $Z$ 取值为 1 的组的均值减去 $Z$ 参照组的均值之差。

三向交互作用系数告诉我们给定的自由主义—保守主义分数变动一个单位,双向交互作用系数变动的数量。三向交互作用是 1.007,这表示自由主义倾向增加一个单位,前述交互作用参数 －17.241 将等于 －17.241 ＋ 1.007 ＝ －16.234。如果自由主义倾向增加两个单位,那么前述交互作用参数将变成:

$$-17.241 + 1.007 + 1.007 = -15.227$$

随着自由主义倾向的增强,被告种族和陪审员种族的交互作用的大小趋近于 0。

正如前面那个例子,选定一些连续性二阶变量的值,展示相关的 2×2 表,这有助于读者理解三向交互作用。表 3.6 给出了自由主义—保守主义倾向的“低”、“中”、“高”值。“低”定义为比自由主义—保守主义分数均值低一个标准差,“中”是均值,“高”则是比均值高一个标准差。使用变换策略和“虚拟变量重新编码”策略来分离出相应的统计数值。如此安排,以便关键自变量由行来表

示，而列是一阶调节变量，二阶调节变量用于分隔这个 2×2 表。每一个 2×2 表中的单元格都是交互作用的参数。三向交互作用由此变得很明显，因为双向交互作用估计值在不同的自由主义—保守主义层次上变动。可用先前讨论的原则进行图表展示。

**表 3.6 双向交互作用对定罪概率影响的均值(在自由主义—保守主义倾向的不同层次上)**

| | | AA 陪审员 | EA 陪审员 |
|---|---|---|---|
| 低自由主义倾向 | AA 被告 | 49.18 | 80.92 |
| | EA 被告 | 50.07 | 48.31 |
| | | (49.18－50.07)－(80.92－48.31)<br>=－17.24** | |
| | | AA 陪审员 | EA 陪审员 |
| 中自由主义倾向 | AA 被告 | 40.23 | 56.24 |
| | EA 被告 | 40.23 | 39 |
| | | (40.23－40.23)－(56.24－39.00)<br>=－0.97 | |
| | | AA 陪审员 | EA 陪审员 |
| 高自由主义倾向 | AA 被告 | 31.27 | 31.56 |
| | EA 被告 | 30.39 | 29.71 |
| | | (31.27－30.39)－(31.56－29.71)<br>=－33.50** | |

注：AA＝美国黑人；EA＝美国白人。
** 表示 $p<0.05$。

### 两个类别以上的定性变量

上述逻辑很容易扩展到有两个以上类别的定性变量的情况。我们必须谨记这些系数的条件性本质以及特定设计中单元格所代表的乘积项。例如，在第一个例子中，关键自变量是连续变量。假设被告的种族有 3 个类别（美国黑人、美国白人、拉美人），并且陪审员也有 3 个类别（美国黑人、美国白人、拉美人），那么全模型就是：

$$
\begin{aligned}
Y = & a + b_1 L + b_2 AC_{AA} + b_3 AC_L + b_4 JU_{AA} + b_5 JU_L \\
& + b_6 (L)(AC_{AA}) + b_7 (L)(AC_L) + b_8 (L)(JU_{AA}) \\
& + b_9 (L)(JU_L) + b_{10} (AC_{AA})(JU_{AA}) + b_{11} (AC_{AA})(JU_L) \\
& + b_{12} (AC_L)(JU_{AA}) + b_{13} (AC_L)(JU_L) \\
& + b_{14} (L)(AC_{AA})(JU_{AA}) + b_{15} (L)(AC_{AA})(JU_L) \\
& + b_{16} (L)(AC_L)(JU_{AA}) + b_{17} (L)(AC_L)(JU_L) + e \qquad [3.2]
\end{aligned}
$$

这里的 $L$ 是对中的自由主义—保守主义分数。$AC_{AA}$ 是一个虚拟变量（被告是美国黑人时赋值为 1，其他为 0），$AC_L$ 也是虚拟变量（被告是拉美人时赋值为 1，其他为 0）；$JU_{AA}$ 是虚拟变量（陪审员是美国黑人时赋值为 1，其他为 0），$JU_L$ 也是虚拟变量（陪审员是拉美人时赋值为 1，其他为 0）。被告种族的参照组是美国白人，而陪审员种族的参照组也是美国白人。尽管这个方程看上去很复杂，但用这种讨论原则其实很好操作。总体三向交互作用通过运用方程 1.1 的层级检验，即对方程 3.2 与去掉 $b_{14}$、$b_{15}$、$b_{16}$ 和 $b_{17}$ 系数相关变量项的模型进行比较。系数 $b_1$ 是美国白人陪审员为美国白人被告定罪时，自由主义—保守主义倾向的作用。重新定义虚拟变量，重新计算方程，并分离出两个调节变量不同的参照组组合，就会产生 9 个斜率值相关统计，由 $3 \times 3$ 被告种族和陪审员种族表定义。在每一种情况下，关注点都是 $b_1$。

任何包含 $L$ 的双向交互项都反映了斜率之间的差异。例如，$b_7$ 代表的是陪审员为美国白人时，定罪分数对自由主义分数进行回归，被告为拉美人的斜率均值减去被告为美国白人的斜率均值的差。一个给定的三向乘积项反映的是斜率之间差异的差。例如，$b_{17}$ 就是美国白人陪审员斜率差异减去拉美人陪审员的斜率差异。尽管涉及很多系数，对它们的解释还是很简单的。

## 小结

研究者常对探索以三向交互作用为焦点的交互模型感兴趣。交互作用中的解释变量可能全是连续的，或它们是连续和定性变量的组合。在这两种情况

下，研究者都宣称第一个变量为关键自变量，第二个变量是一阶调节变量，第三个变量是二阶调节变量。恰当的乘积项由此可被计算出来，然后交互模型包含了所有“主效应”项和所有乘积项。总体交互作用的显著性检验反映在与三向乘积项相连的回归系数里，并具有单一自由度。当总体交互作用超过一个自由度时，就需要运用层级检验比较三向交互模型和去除三向乘积项模型来评估总体的统计显著性。

非标准化交互效应的优势在于，回归系数反映的是单自由度的交互比较。标准化交互效应的优势在于其反映了三向乘积项的半偏相关系数平方。

对交互作用的理解通常关注与关键自变量相关的回归系数以及包含关键自变量各个乘积项的回归系数。针对不同的调节变量，我们使用不同的对中和重新编码策略来帮助读者理解交互的本质。

# 第 4 章 | 其他重要问题

本章考察一系列与交互分析相关的事项，大部分内容都比前面讨论的更高深。

## 连续变量之交互作用的双线性本质

如前所述，当一个连续变量是交互作用的一部分时，重要的一点就是传统的包含乘积项检验的交互模型有一个特定形式，即双线性交互。

其他形式的交互也可操作，并且应当常规性地进行探索性分析，以保证交互作用的形式是正确的。在两个连续变量的情况下，经典乘积项方法反映了一个定义较窄，但适用性可能更宽的交互形式。如前所述，如果 $X$ 是关键自变量，$Z$ 是调节变量，那么乘积项方法会把 $X$ 的系数当做 $Z$ 的一个线性函数，但很可能 $X$ 的系数变动是 $Z$ 的非线性函数。如果那样，传统的乘积项分析就是设定有误的模型。探究这一问题的一个粗略但具有启发性的方法是使用分组差异回归(Hamilton, 1992)。在此情况下，调节变量($Z$)按照差不多相等的数量被分成5到10个组，顺序分类。然后，给计算每一组中 $Z$ 的均值或中位数，再对每一个 $Z$ 组进行一次 $Y$ 对 $X$ 的回归分析。在由 $Z$ 定义的5到10个组中，检查 $Y$ 对 $X$ 的回归系数，这揭示了根据每组 $Z$ 的均值或中位数的变化，$Y$ 对 $X$ 回归系数是否存在增加或者减少的趋势。换言之，如果我们根据 $Y$ 对 $X$ 回归的系数与 $Z$ 的每组均值(或中位数)作图，就会存在一个线性趋势。如果并非这样，那么就需要一个更复杂的交互形式。

这种复杂的交互形式往往可使用乘积项的多项式回归。包含交互项多项

式分析的导论，请参见 Jaccard、Turrisi 和 Wan(1990)。举个例子，$X$ 和 $Z$ 都是连续变量，要应用这样的模型，假设 $X$ 的系数不是 $Z$ 的线性函数而是 $Z$ 的二次函数，步骤如下：(1) 确定关键自变量 $X$ 和调节变量 $Z$；(2) 对 $X$ 和 $Z$ 进行必要的变换(比如对中操作)；(3) 计算调节变量的平方，$Z^2$；(4) 计算 $X$ 和 $Z$ 以及 $X$ 和 $Z^2$ 的乘积项；(5) 确定方程 $Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 Z^2 + \beta_4 XZ + \beta_5 XZ^2 + \varepsilon$。

进行层级检验可观察加入 $XZ^2$ 项对模型的改进程度，表示这个二次项交互作用是否重要(或可简单地检查与 $\beta_5$ 相连系数的显著性)。给定 $Z$ 值，$X$ 的系数是由 $\beta_1 + \beta_4 + \beta_5 Z^2$ 决定的。$\beta_1$ 是当 $Z = 0$ 时，$X$ 的系数。可以变换 $Z$(见第二步)，使 $Z$ 的 0 值代表一个理论上有意义的值，以此来分离在任意给定的 $Z$ 值上，$Y$ 对 $X$ 的系数和置信区间。

对只涉及定性变量和连续解释变量的情况而言，假设 $Z$ 是虚拟变量，取值为 1 和 0，表示属于哪个组。在此情况下，至少在一个组上(很可能在两个组上)，$Y$ 都是 $X$ 的一个非线性函数。拟合下面这个方程：

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 X^2 + \beta_4 XZ + \beta_5 X^2 Z + \varepsilon$$

当 $Z = 0$ 时，$X$ 对 $Y$ 的作用反映在二次模型 $\alpha + \beta_1 X + \beta_3 X^2$ 中。为找出当 $Z=1$ 时 $X$ 对 $Y$ 的作用，必须重新给 $Z$ 编码，使 $Z$ 的参照组调转，重新计算乘积项，重新进行回归，并再次计算 $\alpha + \beta_1 X + \beta_3 X^2$。

有一些方法论学者(例如 Ganzach, 1997)认为，假设简单线性关系和简单双线性交互关系的模型都过于有限，在研究中应当常规性地考察曲线关系效应。方法之一是拟合一个模型，允许 $Y$ 和解释变量之间以及与交互项之间产生线性或二次的关系。同时，该模型会对 $X$ 系数对 $Z$ 的线性或者二次函数形式很敏感。对连续变量而言，这种模型会有如下形式：

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 Z + \beta_4 Z^2 + \beta_5 XZ + \beta_6 XZ^2 + \beta_7 X^2 Z + \beta_8 X^2 Z^2 + \varepsilon \quad [4.1]$$

在任意的给定 $Z$ 值上，$X$ 的作用都可以被分离出来，通过变换 $Z$ 值策略，使 $Z$ 的 0 值等于我们感兴趣的那个值，然后关注 $\beta_1$ 和 $\beta_2$，分离当 $Z = 0$ 时，相应模

型 $Y = \alpha + \beta_1 X + \beta_2 X^2$ 的系数。正如后面将讨论的，这种模型的优势是帮助理论家避免虚假相关或忽略真实的交互作用。批评者认为，这一方法可能过度拟合数据，且违背理论建构中的简洁原则。这两个观点都有道理。在方程4.1中，我们假设的是任何非线性的关系都是二次形式的。

## 计算关键自变量在调节变量取不同值时的系数

在前面的章节中，关键自变量系数都是在调节变量取不同值时计算出来的，即通过转换连续性调节变量或重新定义定性调节变量的参照组，然后在电脑上重新进行回归。

尽管这个方法有些繁琐，但其优势是可以产生所有我们感兴趣变量的标准误和置信区间。手工计算这些置信区间就不那么容易(相关方程见 Aiken&West，1991)。当要计算原方程中的系数而不产生置信区间，不利用变换后的变量重新进行回归时，就会有问题。这一节将描述这么做的基本逻辑。让我们从方程2.7中的方法开始，再把这一逻辑推广到其他情境下。

请考虑 $X$ 是关键自变量，$Z$ 是调节变量的情况，方程是：

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon \qquad [4.2]$$

要决定某些 $Z$ 值上 $X$ 的系数，首先必须分离出右边包含 $X$ 的所有的项：

$$\beta_1 X + \beta_3 XZ$$

然后，我们把 $X$ 提出来：

$$X(\beta_1 + \beta_3 Z)$$

产生的就是在任意给定 $Z$ 值上的 $X$ 值：

$$X \text{在} Z \text{值上面的系数} = \beta_1 + \beta_3 Z \qquad [4.3]$$

在方程4.2中，如果 $\beta_1 = 1.2$，$\beta_3 = 0.05$，那么当 $Z = 2$ 时，$X$ 的系数就是 $1.2 + (0.05)(2) = 1.3$。请注意，当 $Z = 0$ 时，方程4.3中的系数就是 $\beta_1$，这就表明 $\beta_1$ 是以 $Z = 0$ 为条件的。

如果 $X$ 和 $Z$ 都是虚拟变量，那么方程 4.3 的逻辑依然成立，不过关注点就是相关虚拟变量了。例如，假设 $X$ 有两个虚拟变量，而 $Z$ 有两个虚拟变量，就产生了下面的方程：

$$Y = \alpha + \beta_1 D_{X1} + \beta_2 D_{X2} + \beta_3 D_{Z1} + \beta_4 D_{Z2} + \beta_5 D_{X1} D_{Z1}$$
$$+ \beta_6 D_{X1} D_{Z2} + \beta_7 D_{X2} D_{Z1} + \beta_8 D_{X2} D_{Z2} + \varepsilon$$

假设我们想分离出 $D_{Z1}$ 和 $D_{Z2}$ 都为 1 时，在 $X$ 上 $D_{X1}$ 为 1 的组与参照组的系数。首先分离直接包含 $D_{X1}$ 的项和系数：

$$\beta_1 D_{X1} + \beta_5 D_{X1} D_{Z1} + \beta_6 D_{X1} D_{Z2}$$

然后把 $D_{X1}$ 提出来，我们得到：

$$D_{X1}(\beta_1 + \beta_5 D_{Z1} + \beta_6 D_{Z2})$$

因此，

$$X \text{ 在 } D_{Z1} \text{ 和 } DZ_2 \text{ 上的 } \beta = \beta_1 + \beta_5 D_{Z1} + \beta_6 D_{Z2}$$

在这个例子中，$\beta_1 = 0.2$，$\beta_5 = 0.3$，$\beta_6 = 0.4$，$D_{Z1} = 1$ 且 $D_{Z2} = 1$，那么 $D_{X1}$ 的系数就是 $[0.2 + (0.3)(1) + (0.4)(1)] = 0.90$。[10]

三向交互作用使用相同的逻辑。对 3 个连续性解释变量 $X$、$Q$ 和 $Z$ 而言，传统的交互方程是：

$$Y = \alpha + \beta_1 X + \beta_2 Q + \beta_3 Z + \beta_4 XQ + \beta_5 XZ + \beta_6 QZ + \beta_7 XQZ + \varepsilon$$

在给定 $Q$ 值和 $Z$ 值的组合时，$X$ 的系数是：

$$X \text{ 在 } Q \text{ 值和 } Z \text{ 值组合上的 } \beta = \beta_1 + \beta_4 Q + \beta_5 Z + \beta_7 QZ$$ [11]

而在给定 $Z$ 值时，$XQ$ 的系数是：

$$XQ \text{ 在 } Z \text{ 值时 } \beta = \beta_4 + \beta_7 Z$$

## 计算偏组成项

有时人们会说，乘积项在回归方程中代表交互效应，通过它们反映出乘积

项是主效应和交互效应的混合体。只有当乘积项的组成部分与乘积项一起被包括在方程中时,此书描述的顺序关系才成立(与无限制的截距项一起)。排除一个或者更多乘积项的组成部分依然可能建立交互作用模型,但这就不同于我们这里所考虑的交互作用了。

传统的交互分析使用的是我们称之为"层级完整建构"模型。一个层级完整建构模型是指,所有的高阶交互项涉及的低阶组成项都被包含在模型中。例如,我们对 $X$ 和 $Z$ 之间的双向交互作用感兴趣,那么层级完整建构模型就应当包含解释变量 $X$、$Z$ 和 $XZ$;如果我们对 $Q$、$X$ 和 $Z$ 之间的三向交互作用感兴趣,那么层级完整建构模型就应当包括解释变量 $Q$、$X$、$Z$、$QX$、$QZ$、$XZ$ 和 $QXZ$;如果有一个定性解释变量(用虚拟变量 $D_1$ 和 $D_2$ 表示)和一个连续解释变量 $Z$,那么层级完整建构交互模型就应当包括 $D_1$、$D_2$、$Z$、$D_1 \times Z$ 和 $D_2 \times Z$。研究者使用层级完整建构模型时,就可用本书介绍的顺序关系了。

当然,也可以不使用层级完整建构模型,而对特定类型的交互作用建立模型。例如,一个简单的相乘模型可以采用如下形式:

$$Y = \alpha + \beta_1 XZ + \varepsilon$$

这一模型(例如,乘积相关)的拟合受到 $X$ 和 $Z$ 度量的影响。层级完整建构模型使用的是变换,把 $X$ 减去一个常数而不影响模型拟合,这在上面这个相乘模型中就不适用。如果一个度量在纯相乘模型中是随意的,那么这一模型的拟合也是如此。模型化非层级完整建构的交互作用需要熟练的技术才可运用。

## 变换

我们很依赖简单变换策略(测度值减去一个常数)来分离出简单效应以及相关的标准误、显著性检验和置信区间。这些变换可用于双向交互作用、三向交互作用、四向交互作用或更高阶交互作用。分析者简单地使用层级完整建构模型,并牢记系数的如下条件性本质:在任何时候,变量 $X$ 只要涉及乘积项,那么与它相关的简单系数就都是以乘积项中其他变量取 0 值为条件的。与此类

似，如果一个乘积项（$XZ$）涉及更高阶的乘积项（例如，$XZQ$），那么它与 $XZ$ 相关的系数就都以更高阶乘积项中的其他变量取 0 值（$Q=0$）为条件。

调节变量取某个值时，我们试图分离 $X$ 对 $Y$ 的简单效应，此时若使用变换策略，通常以原始分数来操作为最佳选择，即应用变换产生的 0 值对应我们感兴趣的那个值。有些研究者先对中 $Z$ 分数，再应用变换策略来转换这些 $Z$ 分数（减去或加上一个标准差）。尽管可以这么做，但根据经验，应用这种双重变换分数时，研究者往往会得到令人吃惊的结果。用代数来解释其背后的机制显得有点多余，我们在此提醒读者，最安全的方法是避免对解释变量进行双重变换，且总是从理论需要出发确定所需调节变量的值，通过应用变换策略在原始分数上减去一个常数，进而产生一个我们感兴趣的 0 值点。

## 多重交互作用

请考虑这种情况：研究者需要模型化结果变量 $Y$ 作为 3 个连续变量 $X$、$Q$ 和 $Z$ 的函数。研究者不希望看到解释变量之间的三向交互作用，但试图评估所有可能的双向交互作用。在此，可使用多种策略。有些分析者使用组块检验，先拟合一个模型，把所有双向交互项都包含进来，再与一个没有任何交互项的模型进行比较（Kleinbaum，1992）。利用方程 1.1 可进行这一检验。如果两个模型的拟合程度的差异微不足道，那么就说明没有任何一项交互作用是必要的，它们都可从方程中剔除。如果应用组块检验有明显差异，那么表明至少有一项交互项需要被保留。在此意义上，层级逆向消除策略可用于比较包含所有交互项模型与去掉某一特定交互项模型之间拟合程度的差异（通过方程1.1）。例如，研究者对评估 $XZ$ 的交互作用感兴趣，那么就应当比较这一模型的拟合程度：

$$Y=\alpha+\beta_1X+\beta_2Z+\beta_3Q+\beta_4XZ+\beta_5XQ+\beta_6QZ+\varepsilon$$

与下面模型的拟合程度：

$$Y=\alpha+\beta_1X+\beta_2Z+\beta_3Q+\beta_4XZ+\beta_5QZ+\varepsilon$$

如果这两个模型拟合程度差异很小，那就表明 $XZ$ 项可被剔除。然而，如果模型的拟合度差异是显著的，那么这一项就应被保留。

有些分析者用这种方式系统地评估每一个交互项。有些分析者就首先选择一项，如果这项被剔除，再评估余下的项（从方程中剔除已确定的可剔除项）。例如，首先检验 $XZ$，看它是否能够被剔除，并最终决定是否从模型中剔除它。当 $XZ$ 已不出现在模型中时，再评估 $QZ$，关注点是逆向排除法。也就是说，我们评估的是：

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 Q + \beta_4 XQ + \beta_5 QZ + \varepsilon$$

与如下模型相比：

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 Q + \beta_4 XQ + \varepsilon$$

首先基于一些理论上的标准选择可能被排除的项进行评估。有时，我们首先评估全模型中相应系数具有最大 $p$ 值的项，有时则两个标准兼顾。

在多元交互情境下，剔除某些项可以援引很多的模型拟合标准，对相关策略的争论持续存在。深入考察相关情况已经超出本书的范围。感兴趣的读者可参考 Bishop、Feinberg 和 Holland(1975)、Hosmer 和 Lemeshow(1989)以及 Jaccard(1998)的著作。需要提醒读者的是，这些看似“怪异”的情况，在顺序考虑多元交互项时常会发生。例如，组块检验可能提示说至少有一个乘积项应被保留，但对具体每一项的评估却说每一项都应被剔除出模型。或者，单个项的检验证明某一项应被保留，其余所有项都应被剔除，但当其余项被剔除后，原先的这个保留项也变得不显著，只剩下薄弱的解释力。那么，该如何处理这种情况呢？这取决于要回答的理论问题、研究者的统计框架（例如，虚无假设检验、数量估计、置信区间估计）以及数据的形式。在大部分情况下，选择剔除项是直接的、无争议的，但并非所有情况都如此。

当回归方程只包括两个单独的交互项时（比如，3 个连续变量 $Q$、$X$、$Z$ 和乘积项 $XZ$、$QZ$ 都保留在方程中，不包括其他交互项），其理解与前文给出的交互项的系数理解一样，只不过前文中的系数是其他双向交互作用（以及其他的协变量）保持不变的情况下得到的。因此，低阶项系数以所有乘积项涉及的其他

变量取 0 值为条件。

## 标准化与非标准化系数

本书讨论的回归系数通常指与解释变量相关的非标准化回归系数。尽管交互分析中可使用标准化系数，但这些系数可能会使研究者误入歧途，且不符合我们提出的规则。我们一般不建议使用标准化系数，尽管在有些情况下使用它们很正当。举一个标准化回归系数局限性的例子，假设一个简单的双变量回归，即收入对受教育年限进行回归，我们希望能够发现受教育一年的“价值”。这一分析在两个不同的种族群体间展开，即美国黑人和美国白人。假定分析发现，两个群体具有相同的标准化回归系数，表明教育变动一个标准差，收入会变动 0.50 个标准差。研究者或许就此得出结论说，在两个群体中，教育的“价值”是相同的。假设两个群体教育的标准差都是 3.0，但美国白人的收入标准差是 15000，而美国黑人的收入标准差是 6000，那么问题在于，美国白人的非标准化系数是 2500，而美国黑人的非标准化系数是 1000 美元，即对美国白人而言，一年的额外教育预计将增加 2500 美元的收入，而对美国黑人而言，一年的额外教育只增加 1000 美元的收入。两个组之间存在明显的差异，而标准化回归系数不能反映这一点。[12]

标准化分析的问题在于，它对两个群体产生不同的度量单位。因为标准差大小不同，所以标准差这一度量就不同。对美国白人而言，这个度量是 15000美元；对美国黑人而言，这个度量是 6000 美元。在不同的度量上比较这两个群体，就像是以美元为单位度量一个群体的收入，而用英镑度量另一群体的收入。关于标准化系数的其他限制，请参见 Jaccard、Turrisi 和 Wan (1990)的著作。

## 度量性质

度量性质在交互分析中很重要。通过经典的潜变量模型来理解测度问题

是最好的。建构的可观测测度可被看做潜变量的一个指标，代表的是有关变量的真实建构。在实践中，我们不知道某人的真实分数，但可利用可观测测度去估计它。可观测测度不仅受某人真实的潜变量值的影响，还受测量误差的影响。图 4.1 表示这一模型的路径图。如果测量的是可观测测度和潜变量之间的关系，那么图 4.1 代表的是一个回归方程，其中，可观测测度对潜变量进行回归：

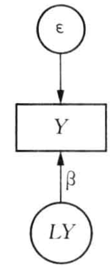


**图 4.1 一个测量模型的路径图**

$$Y = \alpha + \beta LY + \varepsilon$$

上面这个测量模型描述了一个可观测测度如何映射在潜变量上。在这个测量模型上，不同的组在截距、斜率或误差的方差上可能都不同。这些差异会影响在真实的潜变量上群组之间差异的推论(但这却是我们的研究兴趣所在)。例如，两个组在潜变量上具有相同的均值分数，但它们在测量模型中的截距不同，观察变量的平均分数也不同。如果两个组在测量模型中的回归系数也不同，那就说明即使它们在潜变量上有相同的方差，但在观察到的测度上也可能具有不同的方差。如果两个组误差的方差不同，那么测度就依赖于这些组(其他条件保持一致)。最理想的情况是，不同的组由不同的调节变量定义，具有相同的截距和斜率，在测量模型中，误差的方差为 0 或者接近于 0。偏离这一理想状态会导致交互分析的问题。有关度量影响的讨论、检验度量一致性的方法、调整度量的方法，参见 Vandenberg 和 Lance(2000)以及 Busemeyer 和 Jones(1983)的著作。

有一些研究者假定这里讨论的交互分析只对比例尺度适用，但事实并非如此。这种方法可以被有效地运用于定距测度或是与定距测度的特点相当接近的测度。后文会解释这一点。

有一些研究者错误地把“尺度”当做定距或定序的。我们必须认识到，度量的性质与尺度不一致而与数据一致，因此它受数据收集过程中所有方面的影响。一组测度具有定距属性的程度不仅依赖于用于观察的尺度，还受产生观测样本的特定个体、收集时间、数据收集情境等因素的影响。我们来考察一个简单的教学案例，即用两个不同的度量——英寸和高度顺序——来测量 5 个人的身高。

| 个　体 | 高度(英寸) | 身高顺序 |
|---|---|---|
| A | 72″ | 5 |
| B | 71″ | 4 |
| C | 70″ | 3 |
| D | 69″ | 2 |
| E | 67″ | 1 |

众所周知，利用英寸的测度具有定距属性。例如，两个分数之差为 1，对应在身高这个维度上有同样的物理差异。A 和 B 的真实身高差异与 C 和 D 的真实身高差异是相同的，度量的性质反映了这一点(例如，$72-71=1$，$70-69=1$)。同样，D 和 E 的差异是 $69-67=2$，A 和 C 之差异也是 2。这些差异也反映了身高维度上相同的真实差异。请注意，这些属性并不适用于定序测度。A 和 B 的差异是 $1(5-4=1)$，而 D 和 E 的差异也是 $1(2-1=1)$。这个相同的(顺序)差异对应了不同程度的真实身高维度的差异(用英寸进行测度时，D 和 E 之间真实的身高差异明显比 A 和 B 之间真实的身高差异大)。对这些个体而言，定序测度只有定序属性，没有定距属性。

我们再用下面这些分数考察 5 个不同的个体：

| 个　体 | 高度(英寸) | 身高顺序 |
|---|---|---|
| A | 72″ | 5 |
| B | 71″ | 4 |
| C | 70″ | 3 |
| D | 69″ | 2 |
| E | 68″ | 1 |

请注意这 5 个人，定序测度同样具有定距的属性。个体 A 和 B 的差异为 1，这与个体 D 和 E 的差异相同。这些差异所对应的真实物理维度上的差异具有相同的量。在此例中，传统上认为的定序“尺度”事实上产生了一个具有定距属性的测度。假设 E 的身高不是 68 英寸，而是 67.9 英寸。那在此情况下，定序测度在严格意义上就不是定距的。如果它们是没有相反效应的定距层次，那么它们就接近于或可被当做定距来处理。

这个例子揭示了，最关键的问题不在于一组测度是定距还是定序，相反，重要的是一组测度在何种程度上接近定距特征。如果这种相似性非常高，那么就可以假设它们具有定距属性，并利用统计方法对数据进行有效分析；如果这种相似性很低，那么就需要另外的分析策略。在此情况下，交互分析可通过定序层次的数据以及它们所接近的定距层级特征来进行。

## 测量误差

在有关交互作用的统计文献中，大家相当关注一个话题，那就是测量误差带来的偏差。众所周知，在多元回归中，不可靠的测量会产生有偏的回归系数(Bohrnstedt & Carter, 1971)。社会科学家经常使用有误差的测度进行研究。因此，测量误差是交互作用分析的潜在问题，尤其在涉及连续变量时。

Busemeyer 和 Jones(1983)使用经典的检验理论表明，测量误差具有减弱乘积项层级估值的效果。减弱程度是乘积项可靠性的函数，我们称之为 Pr。在标准的统计假定下，减弱的量反映在主效应模型 $R^2$ 的变化上，为 Pr $(R_2^2 - R_1^2)$。例如，如果增加一个乘积项，真实的解释方差增量 $(R_2^2 - R_1^2)$ 为 0.20，而乘积项的可靠性为 0.70，那么，观察到的解释方差将会是 $(0.20)(0.70) = 0.14$，其他条件则保持不变。

在合理的统计条件下[13]，如果 $X$ 和 $Z$ 之间的真实相关性为 0，那么乘积项 $XZ$ 的可靠性就等于 $X$ 的可靠性乘以 $Z$ 的可靠性。因此，若一个测度是相对可靠的 $(r_{XX} = 0.80)$，第二个测度是相对不可靠的 $(r_{ZZ} = 0.50)$，那么乘积项的可靠性就会比可靠性最低的组成项的可靠性还低，即$(0.80)(0.50) = 0.40$。当 $X$ 和 $Z$ 的真实相关性增大时，乘积项的可靠性也会增大。但根据社会科学中通常观察到的相关性和可靠性可知，这并不会增加太多乘积项的可靠性(Busemeyer & Jones, 1985)。这些事实强调了测量误差给检验交互作用带来的困难，特别是在低统计解释力的情况下。对假设检验而言，利用大样本可以抵消测量误差所降低的解释力，但一个大样本不必然消除测量误差导致的回归系数的偏差(例如，Busemeyer & Jones, 1983; Evans, 1985)。

有几个处理测量误差问题的解决方案。Cohen 等人(1975)、Bohrnstedt 和 Marwell(1978)、Heise(1986)、Fuller 和 Hidiroglu(1978)建议的方法都要求预先对组成变量的可靠性有所了解。Cohen 等人(1983: 410)发现,他们在 1977 年的书中提出的修正方法是错误的,那个方法倾向于高估回归系数。Bohrnstedt和 Marwell 的方法有几个局限,具体见 Busemeyer 和 Jones(1983)的著作。Heise(1986)发现,他的方法只在一开始可靠性就相对较高的情况下才能得到令人满意的效果(比如,高于 0.90)。Fuller 和 Hidiroglu 的方法很有前景,但只针对那些不含乘积项的模型。若把这种方法扩展到乘积项进行分析,将会很有价值。

第二组处理测量误差的方法是使用潜类变量结构方程模型(有关这类方法的导论,见 Jaccard & Wan, 1996)。这些策略依赖于每个真实变量建构都有多个指标,把误差理论整合进模型检验和参数估计中。使用此类方法的交互模型受到极大的关注。这类方法的问题是需要大样本,并要假定解释变量都是正态分布的。有几个策略可以放松正态分布假定。这包括由 Bollen(1996)、Bollen 和 Paxton(1998)发展出来的二阶段最小方差方法(2SLS)、准最大似然估计法(Klein & Muthen,未刊稿)以及含误差因子分数法(Wall & Amemiya, 2000)。目前并不能判断哪个分析方案必然优于另一个分析方案。尽管还有许多工作需要做,但这类方法依然是有前景的。

依靠传统多元回归方法的分析者在分析交互作用时,必须承认存在由测量误差引起的潜在偏差。分析者应当使用有效和可靠的度量,并在作结论时保持必要的警惕。忽视测量误差等于假定有完美的可靠性,这意味着社会科学在进行复杂的理论检验前,应花费相当多的时间和精力来发展高质量的测度。心理计量学、问卷设计和精神物理学中存在大量关于降低测量误差建议的文献。对这些实践的有用讨论,请参见 Anderson(1981)和 Wegenar(1982)的著作。

## 稳健分析与假定的违反

到此为止,我们的分析都假定有关 OLS 回归推论的标准成立。我们也预设

不存在掩盖数据根本趋势的极端值。

很不幸，这些 OLS 回归的假定常常被违反，推论检验会相应地被弱化。应对假定被违反的一个方法是使用初步检验来评估这些假定的有效性（比如，正态分布检验或方差异质性检验），当发现假定被违反时，应修改相应的分析策略。许多此类检验都缺乏统计解释力，除非使用大样本，否则通常很难发现非正态分布或方差异质性。另外，传统 OLS 的 $t$ 检验和 $F$ 检验并非是在利用"筛选检验"来决定是否继续分析的情况下发展起来的。应用这种筛选，可能改变样本的内在分布，弊大于利。

另外一个策略就是以某种方式变换数据，使统计模型拟合数据。有时候，这个变换解决了一个问题（非正态），同时产生了其他问题（方差异质性）。有些变换改变了变量单位，使新度量不具备真实世界的意义。从实践角度来看，这使测度问题变得难以处理。最后，通常的变换哲学是理顺本末关系。与其通过变换来操作数据以符合潜在的统计模型，何不直接使用一个适合于已有数据的统计模型呢？

20 多年前，在很多情况下，发现这样一个统计模型或许会被证明是不可能的。但随着电脑技术的迅速发展，稳健统计方法领域有了长足的进步。现在有切实可行的分析方法，并且它们不像传统统计方法那样要求强假定，这就产生了非常好的比较统计力量，且不用回避极端值。有些方法使用了本书讨论的原则（比如，乘积项产生同类型的系数理解），却是在不回避极端值且产生稳健标准误和稳健置信区间的情况下。有些方法涉及完全不同的框架，比如基于平滑化的方法。关于这些稳健方法，请参考 Wilcox（1997、2001）的著作。Wilcox（1997）特别介绍了关于某些形式交互作用的稳健分析。

最新进展是把稳健方法整合到结构方程模型中，同时涵盖违反假定、极端值和测量误差。这些方法包括标准误、置信区间引导估计以及共同关注不需假定分布的估计（Arbuckle & Wothke，1999）。这些技术拥有广阔的前景。

## 个体内测量和重复测量设计

有时,研究者要求在个体内或是重复测量设计情况下,检验斜率之间的差异。设计之一是在两个时间点上测量结果变量,然后对两个时点共同的、稳定的解释变量进行回归。例如,分别用受测试者在七年级时和九年级时与父母关系的满意度对性别进行回归。我们感兴趣的是性别对关系满意度的作用在七年级时是否与在九年级时不同。这种设计产生了下面两个方程:

$$Y_{t1} = \alpha_{t1} + \beta_{t1} X + \varepsilon_{t1}$$

$$Y_{t2} = \alpha_{t2} + \beta_{t2} X + \varepsilon_{t2}$$

这里,$X$ 是随时间不变的解释变量,$Y_{t1}$是在时点 1 测量的结果变量,$Y_{t2}$是在时点 2 测量的结果变量。Judd、Kenny 和 McClelland(2002)表明,如果假定图 4.2 的路径模型成立,那么两个方程的斜率相同的虚无假设检验就可通过传统的 OLS 回归来进行。假设 $Y_{t1}$和 $Y_{t2}$对 $X$ 的差异预测如下:

$$Y_{t1} - Y_{t2} = a + bX + e \qquad [4.4]$$

方程 4.4 中的斜率 $b$ 等于$b_{t1}$和 $b_{t2}$的差异,检验 $b$ 的显著性可以评估虚无假设 $\beta_{t1} = \beta_{t2}$。Judd 等(2002)讨论了将这种检验扩展到多于两个重复测度的情况。

在第二个个体内,设计关注 $X$ 和 $Y$ 都随时间变动的个案,产生的两个方程是:

$$Y_{t1} = \alpha_{t1} + \beta_{t1} X_{t1} + \varepsilon_{t1}$$

$$Y_{t2} = \alpha_{t2} + \beta_{t2} X_{t2} + \varepsilon_{t2}$$

虚无假设 $\beta_{t1} = \beta_{t2}$ 即被称为“序列调节”的检验。James 和 Tetrick(1984)描述了一个 $T^2$ 统计量,基于最小方差回归模型来检验序列调节,假定图 4.3 中的模型成立。

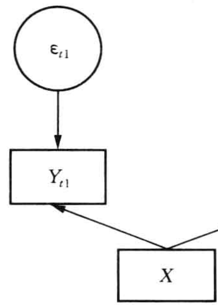



**图 4.2 具有稳定解释变量的纵贯模型**

这些程序的一个局限性是图 4.2 和图 4.3 中的内在模型可能不正确，这会导致检验失效。检验中经常存在相关误差，或者说，在时点 1 上的结果变量可能对时点 2 上的结果变量在 $X$ 之外还有独立的作用。对图 4.3 中的模型而言，$Y_{t2}$ 或许不仅受 $X_{t2}$ 的影响，还受 $X_{t1}$ 独立于 $X_{t2}$ 的影响。如果这个模型是设定错误的，那么这些参数差异的检验就被低估了。检验交互作用假设最好的方法是结构方程模型(SEM)(Kline, 1998)。SEM 可以容纳较多的因果模型和误差结构，在此检验情境下，它们经常要求中等样本或大样本，因为它们依赖于渐进理论。计量经济学中的特定方法对此有所帮助(Greene, 1997)。

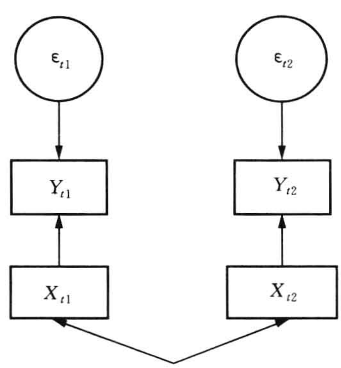



**图 4.3 序列调节模型**

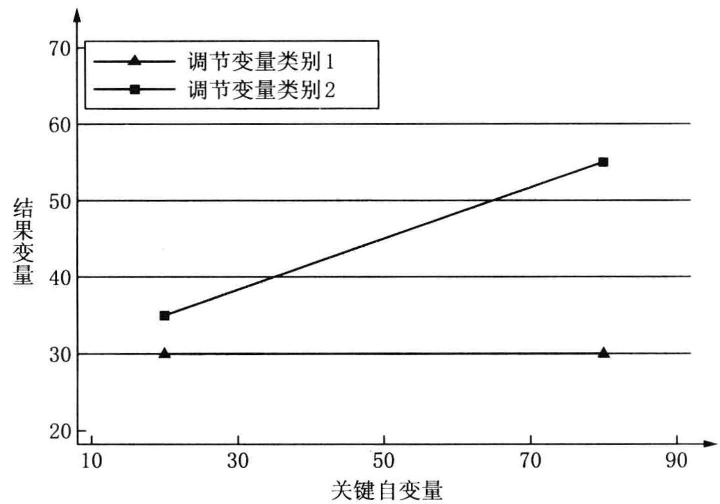
另一个常用的、基于时间的交互分析是增长曲线模型。在此方法中，研究者指定一个数学函数来描述个体的一个变量如何随时间的变化而变化。例如，

某人的阅读能力($Y$)以 1、2、3、4、5 和 6 表示,可能随年级的变动而变动,且这种变动是线性的。可用斜率和截距对这个个体的线性变动进行描述,斜率表示的是个体时间变化一个单位,阅读水平会变化多少个单位。如果时间单位是年,那么这些斜率就表示从这一年到下一年,阅读水平有多少变化(从 1 到 6)。这些斜率可能会有个体差异,有些孩子展示出更陡的斜率,表明从这一年到下一年,阅读水平有更大的变化;另外一些学生则会展示出相对平缓的斜率,表明从这一年到下一年,阅读能力的变化比较小。我们感兴趣的是预测斜率的量级。例如,平均而言,私立学校的孩子相比公立学校的孩子,是否会有更陡的斜率?那个斜率的量级是不是孩子社会经济出身的函数?这些问题(都基于交互作用)可用在层级线性模型统计文献中发展出来的增长曲线模型来解释(Bryk & Raudenbush, 2002)。增长曲线模型可用结构方程模型框架加以参数化(Duncan, Duncan, Strycker, Li & Alpert, 1999)。

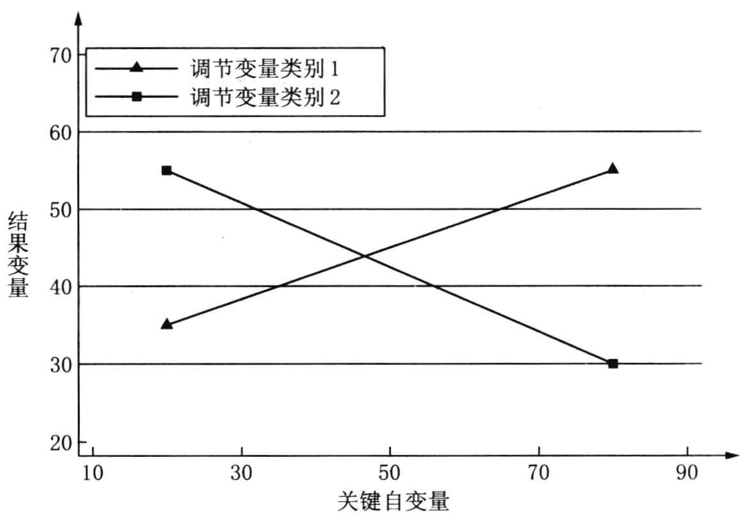
## 同序和非同序交互

社会科学家区分了同序交互作用和非同序交互作用。我们通常在此情境下进行这一区分:一个定性解释变量有二三个组,另外有一个连续性解释变量。非同序交互是指,$Y$ 在一个组中对连续变量进行回归的回归线与 $Y$ 在另一个组中对连续变量进行回归的回归线相交,这也被称为"交叉交互"。同序交互是指两组回归线是非平行的,但并不相交。图 4.4 展示的就是非同序和同序交互作用。

统计学家已经表达了关于交互作用的谨慎态度。他们认为,这种交互可能是因变量度量尺度的加工品。通过 $Y$ 分数的一个单调变换,非平行的回归线常常可以被变成平行的。如果 $Y$ 的度量间距是随意的,那么从科学简洁性的角度来看,执行这种变换以去除虚假的调节作用就是合理的。尽管如此,如果其度量是有意义的,那么同序交互作用就不应被撤销。正如 Cronbach 和 Snow (1981)指出的,这种交互作用从本质上来说很重要,当与成本—收益分析结合时,可对分类决策起关键作用。




(a) 同序交互




(b) 非同序交互

**图 4.4 同序和非同序交互作用**

对任何一组非平行线而言,总存在两线相交的点。在此意义上,理论上来讲,所有交互作用都是非同序的。如果在研究的分数取值范围内(例如,IQ 分数在 90 和 110 之间),回归线不相交,交互作用即被认为是同序的。请考虑两个组的例子,每个组都可被描述为 $Y$ 对 $X$ 回归的线性方程。使用下列方程,可识别

回归线对两个组的相交点是连续变量的哪个值：

$$P_1=(a_1-a_2)/(b_2-b_1)$$

这里，$a_1$ 是第一组结果变量对连续解释变量回归的截距，$a_2$ 是第二组结果变量对连续解释变量回归的截距，$b_1$ 是第一组结果变量对连续性解释变量回归的系数，$b_2$ 是第二组结果变量对连续性解释变量回归的系数。

为了展示相关计算和实际应用，请考虑下面这个例子：一位心理学家对两种不同类型的、针对儿童自尊的疗法的相对作用感兴趣。研究者设计了一个实验，旨在通过干预增强孩子的自尊心。一半孩子接受干预A，另一半孩子接受干预B。分组是随机的。研究没有控制组，因为先前的研究表明，与无干预情况相比，这两种干预方法都能有效地增强孩子的自尊心。孩子对这两种干预的反应被认为受到其与父母关系质量的调节作用。对与父母具有相对积极关系的孩子而言，干预方案A比干预方案B更有效；对与父母关系相对较差的孩子，干预方案B比干预方案A更有效。接受干预的类型采用虚拟编码，使用单一虚拟变量；孩子与父母关系的质量基于一位临床医师的评分，采用一个51分的尺度（从0到50）。这位临床医师与孩子及其父母相处了相当长的时间，并在访谈基础上给出评分，分数越高，关系质量越好。再利用一个0到20的尺度测量自尊心，分数越高，自尊水平越高。在此研究中，干预后的自尊心是结果变量，干预类型是关键自变量，而父母—孩子关系则是调节变量。自尊心分数对虚拟变量、关系质量的测度以及这两个变量的乘积项进行回归（此例未进行任何均值对中操作）。交互作用是统计显著的。

首先，我们计算两个组单独的回归方程，自尊心对关系质量分数进行回归。这可运用第2章中讨论的程序进行分离：

$$\text{干预 A}: Y = 1.193 + 0.098Z + e$$

$$\text{干预 B}: Y = 7.193 + (-0.107Z) + e$$

用一张回归线的图就可表明交互作用是非同序的。两线的相交点是：

$$P_1 = (1.193 - 7.193)/(-0.107 - 0.098) = 29.27$$

此点是 $Z$ 的分数,两组的自尊心分数相同。因此,当关系质量分数等于29.27时,干预后自尊心分数对两个干预条件来说是相同的;当关系质量分数高于 29.27 时,在干预 A 下的干预后的自尊心分数就会比干预 B 下的更高;当关系质量分数低于 29.27 时,干预 B 下的干预后的自尊心分数就会比干预 A 下的更高。

这个分析结果提供了孩子该接受哪种干预方案的建议。如果在干预前,孩子与父母的关系分数低于 29.27,那么他们应当被施以干预 B;如果在干预前,孩子与父母的关系分数高于 29.27,那他们应当被施以干预 A。

Cronbach 和 Gleser(1957)回顾了这些处理的逻辑以及教育、组织和心理学研究中的分类决策。作者提出,受测试者分配到处理组的过程(例如,临床干预、教育课程、工作类型)经常受到非同序交互确定的相交点的指引。那些在相交点右边的人被分到一个处理组中,而交叉点左边的人被分到另一处理组中。相反,同序交互建议对所有人使用同样的处理。

## 显著性的区间

尽管上面的相交点分析很有用,但我们必须承认,在确定相交点过程中存在着样本误差。Potthoff(1964)基于 Johnson 和 Neyman(1936)的经典之作,发展了一套方法,他在非同序交互中建立了一个与交叉点理解相关的"显著性区间"。这一方法关注研究者有一个定性关键自变量、一个连续性调节变量。对两组个案而言,这一方法定义了调节变量的一个取值范围、一个组的成员相比关键自变量其他组的成员,不会有更高的 Y 分数。该方法已经考虑了样本误差(Aiken & West, 1991)。

在前文的例子中,应用下面的描述方法产生了一个值域,即 27.36 到31.07。这表示,当父母—孩子关系质量超出 31.07 时,可以很有信心地说(基于 95%的置信度),接受干预 A 的儿童,其干预后的自尊心会超出那些接受了干预 B 的儿童。当父母—孩子的关系质量低于 27.36 时,同样可以很有信心地预测,接受干预 B 的儿童的干预后的自尊心会超过接受干预 A 的儿童。

相关的值可以通过计算下面这两个值(我们称之为 CV)来获得：

$$CV = [-B \pm (B^2 - AC)^{1/2}]/A \qquad [4.5]$$

为了定义 A、B 和 C 的值，使 $Y$ = 结果变量，$Z$ = 连续性调节变量，$N$ = 包括所有组的关键自变量上有取值的个案总数，$n_1$ = 组 1 的个体数量，$n_2$ = 组 2 的个体数量，$F_\alpha$ = 具有 2 个和 $N-4$ 个自由度的 $F$ 值表，对应于实验者定义的关键 $F$ 值，并与一个预先设定的 $\alpha$ 水平相连(一般是 0.05)，$M_1$ = 组 1 的 $Z$ 均值，$M_2$ = 组 2 的 $Z$ 均值，$S_1$ = 组 1 对 $Z$ 回归的平方和，$S_2$ = 组 2 对 $Z$ 回归的平方和，$E$ = 在组 1 上 $Y$ 对 $Z$ 回归的残差平方和加上在组 2 上 $Y$ 对 $Z$ 回归的残差平方和，$a_1$ = 组 1 上 $Y$ 对 $Z$ 线性回归的截距，$a_2$ = 组 2 上 $Y$ 对 $Z$ 线性回归的截距，$b_1$ = 组 1 上 $Y$ 对 $Z$ 线性回归的斜率，$b_2$ = 组 2 上 $Y$ 对 $Z$ 线性回归的斜率。那么，

$$A = [-2F_\alpha/(N-4)]E[1/S_1 + 1/S_2] + (b_1 - b_2)^2$$

$$B = [2F_\alpha/(N-4)]E[M_1/S_1 + M_2/S_2] + (a_1 - a_2)(b_1 - b_2)$$

$$C = [-2F_\alpha/(N-4)]E[N/(n_1 n_2) + M_1^2/S_1 + M_2^2/S_2] + (a_1 - a_2)^2$$

组内回归方程不应当使用 $Z$ 对中的度量。对这种方法、相关电脑程序以及对更复杂情况的扩展和更加深入的讨论，请参见 Aiken 和 West(1991)的著作。Cronbach 和 Snow(1981)讨论了这些方法的优点和局限性。

## 混合交互

有些分析者提出，交互作用经常与 $X$ 对 $Y$ 的曲线作用混合在一起，因此使交互分析变得复杂(Lubinski & Humphreys, 1990)。数据可能是 $X$ 和 $Y$ 之间曲线关系的一个结果，但当一个交互模型使用 $X$、$Y$ 和另一个变量 $Z$ 来拟合数据时，很可能会得到一个虚假的交互作用。

为了展示基本想法，在此举一个例子，$Y$ 和 $X$ 的内在关系模型是曲线的二次项关系，可用下面的模型表示：

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

而交互模型则表示为：

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

结果是，当 $X$ 和 $Z$ 的相关性增强时，$X^2$ 和 $XZ$ 的相关性也增强了。这意味着，若 $X$ 和 $Z$ 相关，将会在两个模型之间产生混合（Busemeyer & Jones，1983；Lubinski & Humphreys，1990）。针对二次项的本质，内在的关系是曲线的，但拟合一个（线性）交互模型会导致关于交互作用的虚假结论。如果交互和曲线的作用都是可操作的，那么拟合传统交互模型会导致错位交互、虚假的交互或是误导的交互作用，与真实交互作用的正负号都是相反的（Ganzach，1997）。

问题的实质在于拟合了设定错误的模型到数据上，这么做使理论家误入歧途。研究者需要仔细考虑可能适合数据的模型，然后再探索这些相应的模型。如果一个曲线作用在理论上是不可能的或不具备概念合理性的，那就不应被使用。如果一个模型理论上有效，那就应被考虑。可能的情况是，当协变量取曲线作用并代入方程时，交互项就变得统计不显著。这不一定代表交互模型失效了，它只表明存在一个备择模型可解释这些数据。我们的底线是运用常识和理论来组织模型的类型，研究者必须认识到需要考虑多重模型，并在确定交互模型前进行比较。Lubinski 和 Humphreys（1990）及其他人的著作建议研究者在关注交互作用时，考虑曲线模型作为竞争性备择模型的有效性。

## 最佳实验设计和统计解释力

有些分析者哀叹，由于统计解释力低，在线性模型中发现交互作用非常困难。McClelland 和 Judd（1993）探索了大量在实际情境中引发低统计解释力的原因。这些研究者指出，利用统计解释力来发现交互作用，很大程度上依赖于乘积项组成变量的分布本质（例如，$X$ 和 $Z$ 是乘积项 $XZ$ 的组成部分）。McClelland 和 Judd（1993）建议，当实际限制导致小样本和交互作用的低统计解释力时，可对 $X$ 和 $Z$ 的极端个案进行过度取样。然而必须小心地使用这些方法，因为它们会产生有偏的标准化作用大小的估计，例如在比较加叠模型（只含主效

应的模型）和交互模型（全模型）复相关系数平方的增量时。

对交互作用进行统计解释力分析时，要注意交互作用大小的理论限制。例如，对交互项传统的解释力分析基于在“只含主效应”模型中加入一个乘积项，并观察可额外解释多少方差。如果理论表明，在两个解释变量中存在先验的同序交互，且在总体中这种交互作用不为0，那么根据定义，含主效应的总体模型的复相关系数平方必然不为0。同序交互作用越大，主效应模型中的效果也越大，这会被当做基准模型来评估交互项的统计解释力。Rogers（2002）描述了这两个作用大小之间的正式关系。对同序交互作用进行解释力分析时，必须尊重这种关联，因为应用解释力分析时，研究者往往会不经意地选定一个理论上不可能的总体标准化交互作用大小。更多细节请参见 Rogers（2002）的著作。

## 协变量

协变量可被加入到本书讨论的任一回归模型中，这并不改变对相关交互作用的理解。例如，一个协变量 $Q$ 可以被加入到乘积项模型中：

$$Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ + \beta_4 Q + \varepsilon$$

在此模型中，$\beta_3$ 是保持 $Q$ 不变，$Z$ 变化一个单位导致 $X$ 对 $Y$ 作用变动的单位。截距是当 $X$、$Z$ 和 $Q$ 都等于0时，预测的 $Y$ 的均值。对 $Q$ 进行简单线性转换不会影响其他解释变量的系数，但当 $Q$ 的0值点改变后，就会影响截距项。

## 控制实验误差

在有些交互分析中，研究者进行的是多个单自由度比较。出现这种情况时，对一个给定的比较而言，作为标杆的 alpha 保持在一个特定值上（通常为0.05），但至少在一组比较中，发生类型1错误的概率大于标杆 alpha。在此情况下，有些研究者就援引统计调整来维持一个特定的比较水平——实验误差的

alpha水平(例如,在一组比较中至少有一个发生类型1错误的概率)。最流行的方法是传统的Bonferroni程序,尽管这一技术是保守的。更有力的备择方法是有效地运用类型1错误比例(Westfall et. al., 1999)。例如,Holm(1979;亦见Holland & Copenhaver, 1988; Seaman, Levin & Serlin, 1991)建议用一系列的修订版的Bonferroni模型。下面将介绍它们的应用。首先,通过每一组比较,在比较的组中得到一个$p$值。将$p$值从大到小排列。如果$p$值相同,那么它们就随意排序或利用理论标准排列。最小的$p$值与$0.05/k$的alpha值进行比较,$k$是这个组比较的数量。如果这个检验导致拒绝相关虚无假设(因为观察到的$p$值比调整的$\alpha$值小),那么倒数第二小的$p$值就与另一个alpha水平$0.05/(k-1)$进行比较,$k-1$是剩下的比较数量。如果这个检验导致拒绝虚无假设,那么再次小的$p$值就与$0.05/(k-2)$进行比较,如此进行,直到发现差异不显著为止。一旦观察到差异统计不显著,那么剩下的比较都可被视为不显著的。

如果报告的单自由度主要是展示性的,用于向读者提供一个交互作用的含义,那么就不必援引实验控制。如果研究者将基于分析作出理论陈述,此问题就更为重要。援引实验误差控制是复杂的,且会受一系列条件的影响,包括统计解释力和类型1、类型2错误。

## 整体检验和交互作用

交互分析的常见策略是先对交互作用执行整体检验,只有当整体作用统计显著时,才可进行单自由度的交互比较。整体检验作为单自由度交互比较的基础,可以保证研究不会扩大类型1错误。一般而言,大部分二步法都不足以作为控制实验误差比例的方法(Jaccard, 1998; Wilkinson, 1999)。另一个策略就是直接从具有理论意义的单自由度比较入手,独立于整体检验结果来援引控制实验误差的方法(修改的Bonferroni检验)。这并不意味着交互作用的整体检验没有意义。若要反映两个或更多变量的总体交互作用的大小,这种检验就很有用。另外,如果整体交互作用离统计显著性很远,此时就不可能有任何的交互

比较会显著。因此,整体检验是省力的工具。关于此问题的进一步讨论,参见 Jaccard(1998)的著作。

## 常见的误用

在研究文献中,常有一些较差的研究与交互分析相关。我们前面已经提到了一个,它为两个或更多组单独计算回归方程,未对这些差异进行正式检验就宣称组间存在差异。另一个例子是交互分析以中位数为临界点,把一个连续解释变量降格为一个两值指标。这一策略常常被用来使传统方差分析法可进行交互作用分析。这种实践是不可取的,因为它们忽略了有用的信息,常常导致更低的统计解释力,并且它们引入了虚假作用(Maxwell & Delaney, 1993)。运用普通的线性模型可以有效地分析连续变量和定性变量之间的交互作用,而无需求助于降格连续解释变量为粗略的、只有两个值的指标。

## 聚类数据和随机系数模型的交互作用

有些研究设计涉及聚类数据,研究者感兴趣的是聚类效应对 $Y$ 在 $X$ 上的斜率的影响。例如,研究者收集了 5000 个学生的数据(有 50 所学校,每所抽一个 100 学生)。研究者感兴趣的是同侪压力($X$)对吸毒($Y$)的影响以及这种作用是否与学生所在学校的规模($Z$)有关。执行此分析的方法是,首先记录给定个体的 $Z$ 值(学生所在学校的规模,因此来自同一所学校的学生有相同的 $Z$ 值),然后形成乘积项 $XZ$,再进行 OLS 分析,用 $Y$ 对 $X$、$Z$ 和 $XZ$ 进行回归,$N=5000$。这一策略是有问题的。限制之一是总体的残差分布可能不是独立的。来自同一所学校的学生相比其他学校的学生,彼此更为相似,这种机制会导致引入残差的相依性。[14]因此,进行聚类分析必须考虑这一问题,而上面的回归策略并没这么做。

在此情境下,统计学家经常用一个不同于传统 OLS 的统计模型,即随机系数回归或者层级线性回归。比如,我们认为回归系数在不同学校之间的差异可

用下面的方程来模型化：

$$\beta_j = \alpha + \beta Z_j + \varepsilon_j \qquad [4.6]$$

这里的 $\beta_j$ 指的是在学校 $j$ 的 $Y$ 对 $X$ 的回归系数，$Z_j$ 是学校 $j$ 的规模大小，$\alpha$ 是 $\beta_j$ 对 $Z_j$ 回归的截距项，$\beta$ 是 $\beta_j$ 对 $Z_j$ 回归的斜率，$\varepsilon_j$ 是 $\beta_j$ 对 $Z_j$ 回归的残差项。总体而言，$\varepsilon_j$ 被假定为正态分布，均值为 0，在给定 $Z$ 值上有固定的方差。在方程 4.6 中出现 $\varepsilon_j$ 是随机系数回归模型的定义规定的。这样一个模型与传统的伴随乘积项的回归模型不同，方程 4.6 要求用特定的分析模型来估计 $\alpha$ 和 $\beta$ 并检验估计值的显著性。随机系数回归模型本身关注交互作用（因为 $X$ 对 $Y$ 的作用随 $Z$ 的变化而变化），但它们通常应用于聚类数据。对这些方法的讨论，参见 Bryk 和 Raudenbush(2002)的著作。

## 连续与离散解释变量

本书中有些例子使用了多值的、定量的、离散的变量作为解释变量，并且把它们当做连续变量处理。当一个解释变量有多个值并基本符合 OLS 假定时，这个方法是可行的。如果一个离散变量只有很少的值，那么我们或许会考虑使用虚拟变量来代表它。另外，可使用回归之外的方法来模型化离散解释变量的作用(Bollen, 1989; Joreskog & Sorbom, 1993)。

一个相关的问题是，隐含的潜在变量对一个建构而言是连续的，而研究者观察到的测度却是离散的。使用比例尺度时常出现这种情况（例如，比例从 1 到 7，用于反映隐含的自尊心这一建构）。分析者在使用传统回归模型时，如果这些测度有许多值，同时基本符合 OLS 回归假定，那一般不会有什么大麻烦。关于克服这些由测度的粗糙性带来的问题，Bollen(1989)和 Joreskog 以及 Sorbom (1993)的著作中有详细的讨论。

## 调节框架的回顾

调节变量框架提供了在实质层面理解交互作用的有效工具。有些统计学

者反对这种方法，因为它从概念上最小化了交互模型中作用的对称性，而且它的关注点离开了那些交互模型中的系数。这些反对声音自有其道理，但用一种理论上合理的概念建立交互框架所带来的收益抵消了这些反对的声音。非调节变量为基础的交互作用概念化强调非加叠性概念以及分离基于残差化均值的效果(Rosnow & Rosenthal, 1996)。调节方法和后一种方法并不存在哪个更优越的问题，它们是看待数据的不同方法。有些方法论学者(如 Pedhazur, 1997)不仅在统计模型意义上，还根据研究设计中的数据类型及其外在特征来定义交互作用。这些定义是艰涩的，且并不局限于交互分析。在思考乘积项的分析意义时，需要把研究设计的特点也考虑进来。在某些情境下，这些分析可能被误解，反映的是调停而非调节，但对更宽层面的应用而言，本书讨论的框架将被证明是对交互分析进行概念化和操作的有用方法。

# 注释

[1] 标准分数计算公式为 $Z=\frac{X-\mu}{\sigma}$，$X$ 是原始分数，$\mu$ 是总体均值，$\sigma$ 是总体标准差。

[2] 当出现抑制变量时，这种对偏相关的理解就具有误导性(见 Cohen & Cohen, 1983)。

[3] 类型 1 错误即虚无假设为真，但显著性检验拒绝了虚无假设，又称“弃真”。

[4] 此处原文是 $X$。根据上下文，可确定是作者或编辑的失误，应当为 $Y$。

[5] 在只有主效应的模型中，$b_1$ 和 $b_2$ 是控制其他解释变量时 $X$ 和 $Z$ 对 $Y$ 作用的大小，而控制的意义是指其他解释变量都取各自的均值。在包含乘积项的模型中，$b_1$ 和 $b_2$ 若要直接代表 $X$ 和 $Z$ 对 $Y$ 作用的大小，就必须控制其他解释变量，但此时的控制一般不是其他解释变量取均值，而是指其为 0。

[6] 此处原文为 6.0＋1.0×2，应是编排的失误。

[7] 原文是 substantial，据文意应当是 substantially different，即变换前，$XZ$ 和 $X$ 的相关性会很高，而变换后，$XZ$ 和 $X$ 的相关性为 0，两者差异巨大。

[8] 原文为 European American，直译为美国的欧洲人。出于汉语习惯，译成美国白人，下同。

[9] 此句原文是“The fact that they are not suggest that there may be three-way interaction”，即指不存在三向交互作用，而事实上，统计结果和图表都表明存在三向交互作用，因此改正。

[10] 原文如此。根据虚拟变量定义，$D_{Z1}$、$D_{Z2}$ 属于同一变量 $Z$，则可同为 0(代表参照组)，但不能同为 1。所以应当只能计算 $D_{Z1}=1$ 且 $D_{Z2}=0$ 时，或 $D_{Z1}=0$ 且 $D_{Z2}=1$ 时的 $X$ 系数，而并不能计算两者同时为 1 时的 $X$ 系数。

[11] 原文是 $\beta_1+\beta_2 Q+\beta_5 Z+\beta_7 QZ$，根据文意，明显应当是 $\beta_4 Q$ 而非 $\beta_2 Q$，故改正。

[12] 标准化系数与非标准化系数的换算公式为 $b_j'=b_j(S_j/S_y)$，$b_j=b_j'(S_y/S_j)$。因此，0.5×(15000/3.0)＝2500，0.5×(6000/3.0)＝1000。

[13] 假定如下：(1)总体中的残差是正态分布的；(2)潜类连续变量是多变量正态分布的(但潜类乘积项不是)；(3)一个给定的观察分数是真实分数的函数，误差分数符合经典检验理论。

[14] 即残差与自变量相关，导致内生性问题。

# 参考文献


Aiken, L. , and WEST, S. (1991) *Multiple Regression*. Newbury Park, CA: Sage.

Anderson, N. H. (1981) *Methods of Information Integration Theory*. New York: Academic Press.

Arbuckle, J. L. , and Wothke, W. (1999) *AMOS 4.0 User's Guide*. Chicago: Smallwaters.

Berry, W. D. , and Feldman, S. (1985) *Multiple Regression in Practice*. Beverly Hills, CA: Sage.

Bishop, Y. M. , Feinberg, G. , and Holland, P. (1975) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Bohrnstedt, G. W. , and Carter, T. M. (1971) "Robustness in regression analysis," in H. L. Costner(ed. ) *Sociological Methodology*. San Francisco: Jossey-Bass.

Bohrnstedt, G. W. , and Marwell, G. (1978) "The reliability of products of two random variables," in K. F. Schuessler(ed. ) *Sociological Methodology*. San Francisco: Jossey-Bass.

Bollen, K. A. (1989) *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K. A. (1996) "An alternative two stage least squares(2SLS) estimator for latent variable equations." *Psychometrika*, 61:109—121.

Bollen, K. A. , and Paxton, P. (1998) "Interaction of latent variables in structural equation models," *Structural Equation Modeling* 5:267—293.

Bryk, A. , and Raudenbush, S. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage.

Busemeyer, J. R. , and Jones, L. (1983) "Analysis of multiplicative combination rules when the cause variables are measured with error." *Psychological Bulletin* 93:549—562.

Cohen, J. , and Cohen, P. (1975) *Applied Multiple Regression for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. , and Cohen, P. (1983) *Applied Multiple Regression for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. (1987) "Statistical tests for moderator variables: Flaws in analysis recently proposed." *Psychological Bulletin* 102:414—417.

Cronbach, L. , and Gleser, G. C. (1957) *Psychological Tests and Personal Decisions* (2nd *ed.* ). Urbana: University of Illinois Press.

Cronbach, L. J. , and Snow, R. E. (1981) *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. New York: Irvington.

Duncan, T. E. , Duncan, S. E. , Strycker, L. A. , Li, F. , and Alpert, A. (1999) *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications*. Mahwah, NJ: Erlbaum.

Evans, M. T. (1985) "A Monte Carlo study of the effects of correlated variance in moderated multiple regression analysis." *Organizational Behavior and Human Decision Process* 36:305—323.

Friedrich, R. (1982) "In defense of multiplicative terms of multiple regression equations." *American Journal of Political Sciences* 26:797—833.

Fuller, W. A., and Hidiroglu, M. A. (1978) "Regression estimation after correction for attenuation." *Journal of American Statistical Association* 73:99—104.

Ganzach, Y. (1997) "Misleading interaction and curvilinear terms." *Psychological Methods* 2: 235—247.

Greene, W. H. (1997) *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.

Hamilton, L. C. (1992) *Regression with Graphics: A Second Course in Applied Statistics*. Belmont, CA: Brooks-Cole.

Hardy, K. (1993) *Regression with Dummy Variables*. Newbury Park, CA: Sage.

Heise, D. R. (1986) "Estimating non-linear models: Correcting for measurement error." *Sociological Methods of Research* 14:447—472.

Holland, B. S., and Copenhaver, M. (1988) "Improved Bonferroni-type multiple testing procedures." *Psychological Bulletin* 104:145—149.

Holm, S. (1979) "A simple sequentially rejective multiple test procedure." *Scandinavian Journal of Statistics* 6:65—70.

Hosmer, D.W., and Lemeshow, S. (1989) *Applied Logistic Regression*. New York: Wiley.

Jaccard, J. (1998) *Interaction Effects in Factorial Analysis of Variance*. Thousand Oaks, CA: Sage.

Jaccard, J., and Dodge. T. (2003) "Specification of contingent effects in linear models," in M. Hardy and A. Bryman(eds.), *Handbook of Data Analysis*. Thousand Oaks, CA: Sage, in press.

Jaccard, J., Turrisi, R., and Wan, C. (1990) *Interaction Effects in Multiple Regression*. Newbury Park, CA: Sage.

Jaccard, J., and Wan, C. (1996) *LISREL Approaches to Interaction Effects in Multiple Regression*. Thousand Oaks, CA: Sage.

James, L., and Tetrick, L. (1984) "A multivariate test of homogeneity of regression weights for correlated data." *Educational and Psychological Measurement* 44:769—780.

Johnson, P. O., and Neyman, J. (1936) "Tests of certain linear hypothesis and their application to some educational problems." *Statistical Research Memoirs* 1:57—93.

Joreskog, K., and Sorbom, D. (1993) *LISREL Ⅷ*. Chicago: Scientific Software.

Judd, C. M., Kenny, D. A., and McClelland, G. (2002) "Estimating and testing mediation and moderation in within subject designs." *Psychological Methods* 6:115—134.

Kleinbaum, D. G. (1992) *Logistic Regression: A Self Learning Text*. New York: Springer.

Kline, R. B. (1998) *Principles and Practice of structural Equation Modeling*. New York: Guilford Press.

Lewis-Beck, M. S. (1980) *Applied Regression: An Introduction*. Beverly Hilly, CA: Sage.

Lubinski, D., and Humphreys, L. (1990) "Assessing spurious 'moderator effects': Illustrated substantively with the hypothesized('synergistic') relation between spatial and mathematical ability." *Psychological Bulletin* 107:385—393.

Maxwell, S., and Delaney, H. (1993) "Bivariate median splits and spurious statistical significance." *Psychological Bulletin* 113:181—190.

McClelland, G. H., and Judd, C. M. (1993) "Statistical difficulties of detecting interactions and

moderator effects." *Psychological Bulletin* 114:376—390.

Pedhazur, E. (1997) *Multiple Regression in Behavioral Research: Explanation and Prediction*. New York: Harcourt Brace.

Potthoff, R. F. (1964) "On the Johnson-Neyman technique and some extensions thereof." *Psychometrika* 29:241—256.

Rogers, W. M. (2002) "Theoretical and mathematical constraints of interactive regression models." *Organizational Research Methods* 5:212—230.

Rosnow, R. L., and Rosenthal, R. (1996) "Contrasts and interaction effects redux: Five easy pieces." *Psychological Sciences* 7:253—257.

Schroeder, L. D., Sjoquist, D. L., and Stephan, P. (1986) *Understanding Regression Analysis: An Introductory Guide*. Beverly Hills, CA: Sage.

Seaman, M. A., Levin, K. R., and Serlin, R. C. (1991) "New developments in pairwise multiple comparisons: Some powerful and practicable procedures." *Psychological Bulletin* 110: 577—586.

Vandenberg, R. J., and Lance, C. (2000) "A review and synthesis of the measurement in variance literature: Suggestions, practices, and recommendations for organizational research." *Organizational Research Methods* 3:4—69.

Wall, M., and Amemiya, Y. (2000) "Estimation for polynomial structural equation models." *Journal of American Statistical Association* 95:929—940.

Wegnar, B. (1982) *Social Attitudes and Psychological Measurement*. Hillsdale, NJ: Lawrence Erlbaum.

Westfall, P. H., Tobias, R., Rom, D., Wolfinger, R., and Hochberg, R. D. (1999) *Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, NC: SAS Institute.

Wilcox, R. R. (1997) *Introduction to Robust Estimation and Hypothesis Testing*. New York: Academic Press.

Wilcox, R. R. (2001) *Fundamentals of Modern Statistical Methods*. New York: Springer-Verlag.

Wilkinson, L. (1999) "Statistical methods in psychology journals: Guidelines and explanations." *American Psychologist* 54:594—604.

# 译名对照表

| | |
|---|---|
| backward elimination test | 逆向消除法 |
| bidirectional or reciprocal causal relationship | 互为因果关系 |
| bilinear interaction | 双线性交互作用 |
| mean centering | 对中 |
| clustered | 聚类 |
| dynamics | 机制 |
| experimentwise error | 实验误差 |
| factorial design | 析因设计 |
| hierarchical regression | 层级回归 |
| hypothesis | 假设 |
| interaction effects | 交互作用 |
| interval | 定距 |
| level | 类别 |
| metric | 度量 |
| moderated relationship | 调节关系 |
| omnibus interaction test | 整体交互作用检验 |
| overfit | 过度拟合 |
| product term | 乘积项 |
| robust | 稳健 |
| scale | 尺度 |
| squared semipartial correlation | 半偏相关系数平方 |
| simple main effects | 简单主效应 |
| squared multiple correlation | 复相关系数平方($R^2$) |
| two-way interaction | 双向交互作用 |
| three-way interaction | 三向交互作用 |
| three-term equation | 三项目方程 |
| variance heterogeneity | 方差异质性 |

**Understanding Regression Assumptions**

Copyright © 1993 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

This simplified Chinese edition for the People's Republic of China is published by arrangement with SAGE Publications, Inc. © SAGE Publications, Inc. & TRUTH & WISDOM PRESS 2011.

本书版权归 SAGE Publications 所有。由 SAGE Publications 授权翻译出版。

上海市版权局著作权合同登记号:图字 09-2009-547

**Regression Diagnostics**

Copyright © 1991 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

This simplified Chinese edition for the People's Republic of China is published by arrangement with SAGE Publications, Inc. © SAGE Publications, Inc. & TRUTH & WISDOM PRESS 2011.

本书版权归 SAGE Publications 所有。由 SAGE Publications 授权翻译出版。

上海市版权局著作权合同登记号:图字 09-2009-547

**Regression with Dummy Variables**

Copyright © 1993 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

This simplified Chinese edition for the People's Republic of China is published by arrangement with SAGE Publications, Inc. © SAGE Publications, Inc. & TRUTH & WISDOM PRESS 2011.

本书版权归 SAGE Publications 所有。由 SAGE Publications 授权翻译出版。

上海市版权局著作权合同登记号:图字 09-2009-547

**Interaction Effects in Multiple Regression(2nd edition)**

Copyright © 2003 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

This simplified Chinese edition for the People's Republic of China is published by arrangement with SAGE Publications, Inc. © SAGE Publications, Inc. & TRUTH & WISDOM PRESS 2011.

本书版权归 SAGE Publications 所有。由 SAGE Publications 授权翻译出版。

上海市版权局著作权合同登记号:图字 09-2009-548